Signatures of adaptive evolution in the genetic sequences of human pathogenic RNA viruses

Kathryn E. Kistler

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:
Trevor Bedford, Chair
Jesse Bloom
Michael Emerman

Program authorized to offer degree:
Molecular and Cellular Biology

University of Washington

Abstract

Signatures of adaptive evolution in the genetic sequences of human pathogenic RNA viruses

Kathryn E. Kistler

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

Rapid evolution of human pathogenic RNA viruses can undermine our efforts to control infection and transmission. For instance, seasonal influenza undergoes adaptive evolution directed by selection to evade antibodies, resulting in continual antigenic changes and necessitating nearly annual vaccine updates to match the circulating viruses. Additionally, influenza's propensity to undergo adaptive evolution presents another challenge for vaccine production as it adapts to the eggs the vaccine strain is grown in, often altering antigenicity and impacting vaccine effectiveness. Adaptive evolution leaves certain marks on the genome, which can be identified and interpreted through phylogenetic and sequence-based analyses. In this dissertation, I employ a variety of computational techniques to find and interpret these marks in influenza H3N2 and coronaviruses. First, I systematically identify H3N2 mutations that adapt the virus to replication in eggs, show that epistatic interactions between these mutations constrain the adaptive evolution of H3N2, and describe the potential antigenic impact of these egg-adapted mutations. While influenza's adaptive (and particularly antigenic) evolution is widely-appreciated, it is not as well understood which other RNA viruses undergo similar evolution. Thus, I next utilize several complementary methods to show evidence of recurrent adaptive evolution in seasonal coronaviruses that is localized to the viral gene targeted by human antibodies. Finally, I use novel methods to comprehensively scan the genome of SARS-CoV-2 for evidence of adaptive evolution, identify specific adaptive mutations, and show temporal structure to the evolution of SARS-CoV-2 during the first year and a half of the pandemic. Together, the work in this dissertation demonstrates how genetic sequence data can be used to understand the adaptive evolution of human pathogenic RNA viruses, which informs how these viruses can be most effectively controlled.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you to my advisor Dr. Trevor Bedford whose brilliance and motivation has been an inspiration since day one, and whose generous support has kept me going when I was ready to quit. Without your patience and understanding, I would not be writing this thesis. Thank you also to my committee (Drs. Jesse Bloom, Maitreya Dunham, Michael Emerman, and Harmit Malik) whose feedback has been invaluable. Each of you has an excitement for science that is contagious. To my lab mates: thank you for being role models, teachers, and friends. It has been such a pleasure to share ideas, laughs, and beers with you.

Chapter 1

Introduction

1.1 What is adaptive evolution?

When organisms reproduce, cells divide, or viruses replicate, their genetic material must be copied. The molecular machinery that performs this replication is not perfectly faithful and errors are often introduced at random into the newly synthesized genome. These errors, known as mutations, give rise to different alleles (variants of the genetic sequence). Over time, the frequencies of different alleles within a population will change, resulting in the evolution of the genome.

The change in allele frequencies can be driven by two broad modes of evolution: neutral or adaptive. Neutral evolution occurs when mutations do not alter the fitness of the organism, and thus, in the absence selective pressures. Under neutral evolution, some alleles will be lost at random, causing others to increase in frequency within the population by a stochastic process called genetic drift. The neutral theory of evolution is often thought of as a null hypothesis, with the alternative being adaptive evolution. Adaptive evolution occurs in the presence of positive selection to retain mutations that increase the fitness of the organism and purifying selection to purge deleterious mutations. Adaptive evolution is directional (nonrandom) evolution toward higher fitness driven by selective pressures on random mutations.

Fitness is a measure of the number of offspring produced by an organism. It is an organismal trait that results from phenotype, which is dependent on genotype. The mapping of genotype to phenotype and phenotype to fitness are both dependent on environment. This means that fitness is highly dependent on the environment, and therefore so is selection. For instance, a mutation that allows influenza to enter a human cell more efficiently would increase the fitness of a human influenza virus, but not of an avian virus. The mapping

of phenotype to fitness is complex not only because it is environment-dependent, but also because phenotype is a multidimensional space that affects fitness. For instance, a human influenza virus's fitness can be increased not only by enhanced host cell entry, but also by, for instance, increased viral replication, or improved immune evasion.

The complex interactions between genotype, phenotype and fitness that directs adaptive evolution can be visualized as a fitness landscape (Wright, 1932) where the height of the three-dimensional surface represents fitness and the xy-coordinates represent different genotypes (Figure 1.1). Diffusion along the xy-plane of the landscape is achieved by mutation, and uphill movement is driven by selection.

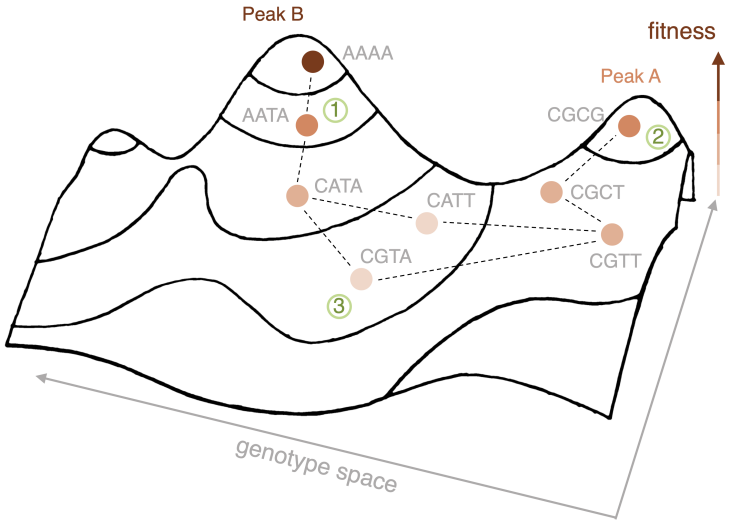

Figure 1.1: Adaptive evolution can be visualized by a fitness landscape. The mutational path from one fitness peak to another is shown on a fitness landscape of a hypothetical genetic sequence. Each genotype along this path is marked by a circle according to its position within the genotype space (xy-plane) and colored according to fitness (z-axis). In this hypothetical scenario, the genomes at points 1 and 2 have the same fitness. However, the genome at point 1 can easily access a higher fitness state through selection on an A $\rightarrow$ T mutation. For the genome at point 2 to access this higher fitness state, it will have to pass through a valley of lower fitness (point 3) caused by reciprocal sign epistasis between genetic positions 2 and 4. While the genome at point 1 clearly has adaptive potential, the fitness valley may be too great for the genome at point 2 to evolve adaptively.

The fitness landscape illustrates how adaptive evolution can be constrained by the mutational accessibility of higher fitness genotypes. For instance, if genome exists on one fitness peak, which is separated from higher fitness peak by a valley of low fitness and traversal from one peak to another requires many mutations, it will be evolutionarily difficult for the

genome to access that higher fitness peak. A valley between peaks can represent genetic change that improves one trait, at the cost of another (Watabe and Kishino, 2010). For example, a particular mutation to influenza's hemagglutinin gene may allow the virus to escape antibodies, but greatly reduce receptor-binding affinity and, thus, cell entry. However, valleys will exist even on a fitness landscape that maps only the fitness effect of one particular trait, such as receptor-binding. This is due to an additional selective pressure that constrains adaptive evolution in a genetic-background-dependent manner: epistasis.

Epistasis is the dependence of the function of a mutation (or genotype at a particular locus) upon the genotype at another locus, and the topology of the fitness landscape is highly-dependent on it. In the absence of epistasis, the fitness of an organism is a sum of the individual fitness effects of each genetic locus. Epistasis breaks this additivity and influences both the ruggedness (aka peakedness) of fitness landscapes and the altitude of the peaks.

1.2 Why does it matter if viruses evolve adaptively?

Viruses, and particularly viruses with an RNA genome, exhibit extremely high rates of mutation. Typically, RNA viruses incur 10^{-4} to 10^{-5} errors per nucleotide per replication cycle (Holland et al., 1992). This high mutation rate supplies the raw material for rapid evolution. An adaptively-evolving RNA virus can change host preference (Longdon et al., 2014; Wu et al., 2019), increase transmission efficiency (Theys et al., 2018), or optimize immune evasion in the span of months to years (Su et al., 2015; Smith et al., 2004). Therefore, whether or not a virus undergoes adaptive evolution impacts its strategy for infection and replication and is highly relevant when considering interventions against human pathogenic viruses.

A virus's propensity to evolve adaptively contributes to its ability to jump host species and transmit effectively in the new host. This is because viruses reside at a fitness peak in their original host species, and will almost always suffer a large loss in fitness when entering the environment of a new host species. A greater level of standing genetic variation and faster

adaptation enable the virus to increase replication and transmission efficiency in the new host. Thus, cross-species tranmission is more prevalent in rapidly-evolving viruses (Parrish et al., 2008). So, understanding a virus's adaptive potential enables us to direct surveillance efforts towards viruses with higher spillover potential. Importantly, early detection of a spillover event can prevent pandemics by containing the virus before it has had time to adapt to humans.

Viral adaptive evolution can also undermine anti-viral drugs and vaccines. For example, a single substitution in influenza H3N2, confers resistance to the anti-viral drug adamantane, arose in the late-1990s. In under 10 years, this mutation swept through the global population of H3N2 viruses, rendering this drug ineffective against all circulating strains by 2006 (Hurt, 2014).

Whereas evasion of a particular drug results in a single evolutionary sweep that fixes a particular mutation or mutations, the propensity of influenza H3N2 to evolve adaptively means that it also regularly fixes substitutions that escape antibody-mediated immunity. The humoral immune system is also constantly evolving to recognize new pathogens and new epitopes on previously-encountered pathogens. This creates an evolutionary arms race between virus and host that results in recurring adaptive evolution to evade antibodies- a process called antigenic drift. Antigenic drift is a prominent feature of H3N2 evolution (Smith et al., 2004), and because of this, immunity from natural infection and vaccination is short-lasting. A study of roughly 150 people, representing a wide range of ages and influenza vaccination statuses, found that the average time between H3N2 infections is 5 years (Kucharski et al., 2015).

In contrast, measles virus is antigenically stable (Fulton et al., 2015), and immunity against this virus is lifelong. In measles, the lack of adaptive evolution to escape antibody-mediated immunity is due to a large evolutionary barrier. To escape human polyclonal sera, measles virus must acquire mutations in at least five of its immunodominant antigenic sites (Muñoz-Alía et al., 2021). Each of these mutations has the potential to negatively impact

protein function, and an increase in fitness is not conferred by any fewer than 5 mutations. In comparison, a single amino acid change is often sufficient for H3N2 to escape polyclonal sera (Koel et al., 2013; Lee et al., 2019). Therefore, antigenic escape is much more evolutionarily accessible to H3N2 than to measles.

The practical consequence of this is that original measles vaccine, which is based on a strain isolated in 1954, is still fully protective against measles viruses circulating today. Meanwhile, the H3N2 vaccine is updated nearly every year by the World Health Organization (WHO) to match antigenic differences in circulating viruses. Thus, the design of effective vaccines depends on whether or not the virus undergoes adaptive evolution, particularly at sites targeted by neutralizing antibodies.

1.3 How can adaptive evolution be identified?

An adaptively-evolving virus will accrue changes over time that confer higher fitness. Thus, a straight-forward way to identify adaptive evolution is to measure the fitness of a temporal series of viral isolates. However, experimental investigation of adaptive evolution is limited by an *a priori* knowledge of the selection pressure and the ability to test fitness with regard to this pressure in an assay that closely mimics natural evolution. For example, the fitness of influenza variants are often measured *in vitro* by competition assays where the virus is grown in cell culture in the presense of polyclonal sera. Higher fitness in these assays is not perfectly analogous to better replication and trasmission in human hosts. Therefore, despite the utility of experiments that link genotype to phenotype to fitness, in many situations, these experiments are not practical or possible and, instead, adaptive evolution can be identified by the marks it leaves on the genomes of a population over time.

Mutations supply genetic variation to a population and, under neutral evolution, stochastic differences in the number of offspring of each variant results in the fixation or loss of variation by genetic drift. The opposing forces of mutation and genetic drift keep genetic diversity at an equilibrium level, which is dependent on mutation rate and population size.

Without selection, each variant has a probability of fixation equal to its frequency in the population. However, when a mutation is beneficial, the variant will sweep to fixation, reducing the genetic diversity in the population. After the sweep, the population will have more low-frequency mutations than shared diversity. Recurrent selective sweeps will produce a ladder-like phylogeny with periodic collapses of genetic diversity followed increases toward the equilibrium level. The fixed beneficial mutations driving these sweeps occur on the trunk, rather than the tips, of the phylogeny.

Ongoing adaptive evolution also means that positively-selected will accumulate in the population over time. This is because a positively-selected mutation that sweeps through the population will be present in every individual after the sweep. Thus, subsequent beneficial mutations will occur on top of previous ones. If it is assumed that mutations are beneficial when they change the encoded protein in a way that increases fitness, then the adaptive mutations must be nonsynonymous. Thus, adaptive evolution will also result in a surplus of nonsynonymous divergence compared to neutrality.

Many established methods based in population genetics use the above ideas to identify adaptive evolution as particular aberrations from neutral evolution. Sequence and phylogenetic-based methods to identify adaptive evolution by its characteristic impacts on genetic diversity, tree topology, and nonsynonymous fixation include d_N/d_S, McDonald-Kreitman, Tajima's D, and Time to Most Recent Common Ancestor (TMRCA).

1.3.1 d_N/d_S

Selective pressures on a genetic sequence are commonly evaluated by the ratio of nonsynonymous to synonymous divergence d_N/d_S. Nonsynonymous divergence is defined as the number of pairwise nonsynonymous differences observed between the sequences in an alignment, normalized by the number of possible nonsynonymous mutations within that sequence. This metric assumes that synonymous mutations are evolutionarily neutral, while nonsynonymous mutations could be beneficial, neutral, or deleterious. Given this assumption, the

ratio d_N/d_S gives the relative enrichment (or depletion) of nonsynonymous change compared to a neutral expectation. If every nonsynonymous mutation is selectively equivalent to synonymous mutations, $d_{\mathrm{N}}/d_{\mathrm{S}} = 1$. Thus, $d_{\mathrm{N}}/d_{\mathrm{S}} > 1$ suggests that at least some substitutions have been positively-selected. In reality, even adaptively-evolving genes are under functional constraints and deleterious mutations will often push d_N/d_S lower than 1. Because of this $d_{\mathrm{N}}/d_{\mathrm{S}} > 1$ is a fairly strict test for adaptive evolution that does best at identifying recurring positive selection. In the context of RNA viruses, which have high mutation rates, the d_N/d_S ratio can be skewed by violation of the infinite sites model, an assumption that there is never more than one alternate allele at each nucleotide site (Kryazhimskiy and Plotkin, 2008). Additionally, the interpretation of $d_{\mathrm{N}}/d_{\mathrm{S}}$ can be complicated for RNA viruses where the assumption that all synonymous mutations are evolutionarily-neutral is violated by conserved RNA secondary structure (Sanjuán and Bordería, 2011; Witteveldt et al., 2014) and overlapping reading frames (Smyth et al., 2018).

1.3.2 McDonald-Kreitman

The McDonald-Kreitman test compares the genetic variation within a population to the variation between that population and a closely-related outgroup. Specifically, the test counts the number of synonymous (P_{S}) and nonsynonymous (P_{N}) polymorphisms that exist within a population and the number of fixed synonymous (D_{S}) and nonsynonymous (D_{N}) differences that exist between that population and an outgroup. Under neutrality, the ratio of $\frac{P_{\mathrm{N}}}{P_{\mathrm{S}}}$ should equal the ratio of the neutral mutation rate at nonsynonymous sites (μ_{N}) to synonymous sites (μ_{S}), and also equal $\frac{D_{\mathrm{N}}}{D_{\mathrm{S}}}$ (McDonald and Kreitman, 1991). Typically, a 2x2 contingency table is constructed from the counts of D_{N}, D_{S}, P_{N}, and P_{S} and a test of independence is employed to assign a p-value. A significant p-value is indicative of adaptive evolution. An extension of this method calculates the number of adaptive substitutions, and can be used to compute a rate of adaptation (Smith and Eyre-Walker, 2002).

Traditionally, the McDonald-Kreitman test counts polymorphisms within a species, and

uses a closely-related species (or multiple species) as an outgroup to determine fixed differences. However, for rapidly-evolving species, such as RNA viruses, fixations can be determined in relation to a prior time point of the same species (Williamson, 2003). Using serial samples from individuals infected with HIV-1, Williamson (2003) uses this tweak to the McDonald-Kreitman test to show rates of within-host adaptation in the HIV-1 *env* gene. The method described in Bhatt et al. (2011) (referred to as the "Bhatt method" in this disseratation) further optimizes the McDonald-Kreitman test for RNA viruses populations that have been serially-sampled over time (Bhatt et al., 2010, 2011). Rather than assuming all polymorphisms are neutral (Smith and Eyre-Walker, 2002), this method allows a category of high-frequency adaptive polymorphisms (Williamson, 2003), and a category of low-frequency deleterious polymorphisms (Bhatt et al., 2010). Additionally, the Bhatt method employs a proportional counting algorithm to deal with sites that are 2-,3-, or 4-state polymorphic due to high mutation rates of RNA viruses. This reduces the number of Type I errors caused by the infinite site assumption of the original McDonald-Kreitman test. Using this proportional counting method to assign each site to a site category, the number of adaptive substitutions (a) can be estimated by:

$$a = n_{\mathrm{f}} - n_{\mathrm{f}}\frac{n_{\mathrm{m}}}{s_{\mathrm{m}}} + n_{\mathrm{h}} - n_{\mathrm{h}}\frac{n_{\mathrm{m}}}{s_{\mathrm{m}}}$$

where n_{f} and s_{f} are the numbers of nonsynonymous and synonymous fixations, n_{h} and s_{h} are the number of nonsynonymous and synonymous high-frequency polymorphisms, and n_{m} and s_{m} are the number of mid-frequency nonsynonymous and synonymous polymorphisms.

1.3.3 Tajima's D

The Tajima's D statistic tests for positive selection by comparing allele frequencies within a population to the neutral expectation. Under neutrality, the population mutation parameter θ ($\theta = 2N\mu$, where N is population size and μ is mutation rate) can be estimated by the mean number of pairwise differences between sequences (θ_{k}) or the number of segregating sites (θ_{s}). In a population with allele frequency distribution matching neutral expectation,

θ_k=θ_s, and the Tajima's D statistic will equal 0. The Tajima's D statistic (Tajima, 1989) is defined as

$$\frac{\theta_k - \theta_s}{\sqrt{Var(\theta_k - \theta_s)}}$$

Selective sweeps resulting from positive selection on a beneficial mutation alter the topology of a tree, resulting in a star-like radiation of terminal branches after the sweep. This increases the amount of low-frequency variation present at the tips, and thus θ_s, but will not affect θ_k. Selective sweeps can, therefore, be identified by negative Tajima's D statistic. However, a negative Tajima's D can also result from an increase in population size, which also results in elevated θ_s. For RNA viruses, this can cause errors in interpreting the Tajima's D statistic when population size expands during an epidemic (Simonsen et al., 1995). Because this leads to high levels of type I errors, Tajima's D is rarely used to identify adaptive evolution in RNA viruses (Bhatt et al., 2010).

1.3.4 TMRCA

As with Tajima's D, the implementation of Time to Most Recent Common Ancestor (TMRCA) analyses for detection of adaptive evolution is based on the fact that strong directional selection skews the shape of phylogenies (Volz et al., 2013). The phylogeny of a neutrally-evolving sequence will be bushy, with deep branches and long TMRCAs (Bedford et al., 2011). Repeated selective sweeps will cause the tree to adopt a ladder-like shape where the rungs are formed by viral diversification and each step is created by the appearance of a new, fitter variant that replaces previous variants. All sequences along each rung share a common ancestor at the last selective sweep. Thus, selection can be quantified by the timescale of population turnover as measured by TMRCA, with the expectation that stronger selection will result in more frequent steps and therefore a smaller TMRCA measure (Bedford et al., 2011).

1.4 About this dissertation

These phylogenetic and sequence-based patterns are fairly universal of adaptive evolution and the methods discussed above can be applied to a wide variety of organisms. However, implementing these methods often requires tweaks that are specific to the system being studied. For instance, high mutation rates and population growth are typical of RNA viruses, and the ways in which they distort measures of adaptive evolution has been mentioned above. These methods are also fairly insensitive to adaptation on short evolutionary timescales.

In this dissertation, I describe my work to employ and expand on the methods described above in order to identify and characterize adaptive evolution in RNA viruses, in contexts that are relevant to human health. Specifically, I focus on influenza virus and coronavirus. Chapter 2 investigates the adaptive evolution of influenza H3N2 during vaccine production. Most influenza vaccines doses are manufactured by passaging virus in chicken eggs, creating a selective pressure for viral mutations that enhance replication within eggs. In Chapter 2, I use phylogenetic methods to pinpoint specific egg-adaptive mutations, identify epistatic interactions between these mutations that shape the fitness landscape, and estimate the phenotypic consequences of these mutations on vaccine efficacy. During my efforts to follow up on these results through experimental evolution of H3N2 in chicken eggs, the SARS-CoV-2 pandemic began. This spurred a pressing new interest in understanding how SARS-CoV-2 might evolve. Chapter 3 describes work I began in the initial months of the pandemic, focusing on understanding adaptive evolution in other, related coronaviruses that are endemic in humans. In Chapter 3, I employ a variety of available analyses to address whether seasonal coronaviruses evolve adaptively, with a particular focus on adaptive evolution in S1, the primary target of neutralizing antibodies. The work described in Chapter 4 focuses on a period roughly a year and a half after the start of the pandemic. At this time there are experimental indications that SARS-CoV-2 is evolving adaptively, but the aforementioned methods will largely fail to show this. Chapter 4 describes a new method for identifying adaptive evolution that is possible during situations of high genomic surveillance. I use this method to

identify regions of the SARS-CoV-2 genome undergoing adaptive evolution. Additionally, I characterize the temporal dynamics of the adaptive evolution of SARS-CoV-2, compare the pace of this evolution to influenza H3N2, and identify specific adaptive mutations.

Chapter 2

Epistasis and the antigenic impact of adaptation to egg-culturing in influenza H3N2 viruses

2.1 Introduction

Seasonal influenza viruses infect millions of people annually, resulting in hundreds of thousands of deaths globally (World Health Organization). Transmission of this highly infectious disease is primarily curbed by vaccination. Because there are multiple types and subtypes of seasonal influenza viruses, the vaccine is multivalent, typically containing antigens against two type A influenza viruses and at least one type B virus (World Health Organization).

However, seasonal influenza viruses undergo antigenic drift: evolution to escape human antibody-mediated immunity. To be effective, vaccines must be antigenically matched to the diversity of influenza strains currently circulating in humans. This means that the precise formulation of the vaccine must be updated nearly every year to combat rapid antigenic drift of the virus. The nearly annual revision of the vaccine requires an efficient pipeline to design, manufacture, and distribute millions of vaccine doses on a relatively fast timescale.

Prior to each flu season, the World Health Organization (WHO) recommends which strains should be included as components of the next vaccine. These components are then mass-produced, predominantly by growing large amounts of vaccine viruses in chicken eggs. To promote optimal viral growth, the viral surface proteins hemagglutinin (HA) and neuraminidase (NA) from the vaccine strain are reassorted into an egg-adapted background containing the other 6 influenza segments (Harding and Heaton, 2018; Yamayoshi and Kawaoka, 2019). HA and NA are the primary targets of the immune system and, consequently, the site

of most antigenic change. The reassorted virus, known as a candidate vaccine virus (CVV), is then injected into the allantoic cavity of a chicken egg where it replicates for several days before the virus is harvested and inactivated to create the vaccine (Brauer and Chen, 2015).

This method was created in the 1970s (Barberis et al., 2016) and has the advantage of being a validated protocol approved by the proper regulatory bodies with existing infrastructure to churn out a large amount of virus relatively cheaply. However, this method also has disadvantages and chief among these is the accumulation of egg-adapted mutations during vaccine production.

The high mutation rate of influenza viruses, which allows them to rapidly evolve to evade human antibodies (thus necessitating frequent vaccine reformulation), also provides a substrate for selective pressures to act upon during egg-based vaccine production. These pressures can result in the fixation of mutations that are advantageous for influenza replication in chicken eggs, resulting in mass-produced vaccine that that no longer look like the vaccine strains chosen by the WHO. Egg-adapted mutations commonly change the receptor-binding specificity of HA from optimally binding sialic acid on the surface of human cells to more efficiently binding, and thus infecting, the different structure of chicken sialic acid (Harding and Heaton, 2018).

Egg-adapted mutations do not necessarily alter antigenicity, but if they do, they create a serious issue by lowering vaccine effectiveness (VE). A meta-analysis of over 50 individual reports published between 1990 and 2013 indicated that that the VE of the H3N2 component has been consistently and significantly lower than the H1N1pdm2009, H1N1(pre-2009) or type B vaccine components (Belongia et al., 2016). This is thought to be partially due to mutations that occur in the vaccine strains during egg-based production. For instance, H3N2 VE was particularly low in 2016-2017 season when egg-passaging introduced an HA T160K mutation, causing the mass-produced vaccine to differ from the recommended vaccine and the circulating H3N2 strains. A threonine (T) at position 160 is glycosylated while lysine (K) is not, resulting in a substantial change in antigenicity. As a consequence, antibodies

elicited in response to the egg-passaged H3N2 2016-2017 vaccine with 160K were poorly protective against circulating H3N2 viruses with 160T (Zost et al., 2017). The 2003-2004 vaccine also displayed low effectiveness against circulating H3N2 viruses. Though the vaccine and circulating strains differed by 13 amino acids in HA, the antigenic mismatch was shown to stem from just two of these residues, both of which enhance viral growth in chicken eggs (Lu et al., 2005). In 2012-2013, the H3N2 vaccine component also had low VE accompanied by egg-adapted mutations (HA H156Q, G186V, S219Y), though it debated to what extent the low vaccine efficacy was indeed due to antigenic changes caused by these mutations (Cobey et al., 2018; Skowronski et al., 2014; Skowronski and De Serres, 2018). It is difficult to unambiguously separate the contribution of egg-adapted mutations from other factors, such as original antigenic sin, that contribute to low VE. Nevertheless, it is clear that the H3N2 vaccine component frequently acquires mutations during egg-passaging and that these mutations can alter antigenicity and, thus, efficacy of the vaccine.

Though egg-adapted mutations can be avoided by using an egg-free method for vaccine production, it is currently logistically difficult to mass produce the required quantity of vaccine via other methods. Only starting in the fall of 2016 did the United States FDA approve manufacturing mammalian cell-based vaccines that were not previously adapted to grow in eggs (Harding and Heaton, 2018). There are currently two influenza vaccines available in the United States that are manufactured without eggs: Flublok and Flucelvax (Centers for Disease Control and Prevention, 2021). However, as of October 2018, 85-90% of influenza vaccine doses in the US were still grown in chicken eggs (Barr et al., 2018) and, given this, it would be prudent to predict how egg-passaging will change a vaccine genetically and antigenically.

Numerous studies have retroactively described how specific egg-adapted mutations alter the antigenicity of egg-passaged H3N2 vaccine strains. Importantly, these reports provide direct correlations between genetic and antigenic changes, often detailing how structural modifications alter antibody binding (Zost et al., 2017; Widjaja et al., 2006). However,

these results are confined to specific mutations in specific strains and cannot explain how any given H3N2 strain might mutate during egg-passaging. Ideally egg-adapted mutations that lower VE could be avoided by predicting mutations in every candidate vaccine strain prior to vaccine production.

Since the first documented human infection in 1968, tens of thousands of H3N2 viruses have been sequenced, including hundreds of egg-passaged strains (Shu and McCauley, 2017). This wealth of genetic data presents the opportunity to identify broader trends in the occurrence and consequence of egg-adapted mutations via phylogenetic methods. Phylogenetic relationships can be used to infer mutations in egg-passaged strains for which the sequence prior to egg-passaging is not available, allowing a greater number of egg-passaged sequences to be considered. Additionally, phylogenies can reveal patterns across a broader range of H3N2 viral diversity, indicating genetic background specificity of egg-adapted mutations, epistatic interactions between egg-adapted mutations, and consistent antigenic effects associated with specific mutations.

Here, we use phylogenetic relationships to identify genetic predictors of egg-adapted mutations. We show that the two most prominent egg-adapted mutations are mutually exclusive and each have additional epistatic interactions with other mutations, resulting in "mutational pathways" that adapt H3N2 to replication in eggs. We describe the phenotypic differences between these pathways and suggest that the effect of egg-adapted mutations on VE can be minimized by selecting a candidate vaccine virus that preferences one of these pathways.

2.2 Results

2.2.1 Phylogenetic method to infer egg-adapted mutations

Egg-adapted mutations are easily detected in viral strains that were sequenced before and after egg-passaging. However, hundreds of egg-passaged H3N2 strains were not sequenced prior to egg-passaging, making the direct identification of egg-passaged mutations impossible.

To overcome this, we have designed a method to infer egg-adapted mutations based on phylogenetic relationships.

Egg-passaged viruses make up several hundred of the tens of thousands of publicly accessible sequenced H3N2 strains. These sequences cover a vast amount of viral diversity representing genetic divergence from a common ancestor. While the genetic sequence of a single strain cannot, alone, indicate whether a mutation occurred in this strain, a phylogenetic tree built from hundreds or thousands of these sequences can. Phylogenies place each strain in an evolutionary context, revealing which portions of the genome were inherited from an ancestral virus and which are novel mutations.

We constructed a time-resolved phylogeny of H3N2 viruses built from the HA sequences of 570 egg-passaged strains and 2481 other strains that were isolated between 2002 and 2019. The non-egg-passaged (NE-passaged) strains include both unpassaged patient isolates and viruses passaged in mammalian cell culture. All available strains that were sequenced prior to egg-passaging were included in the phylogeny. The other NE-passaged strains were chosen by subsampling all viral sequences in the database to select strains isolated over an even distribution of time and geographic locations.

The resulting phylogeny positions each sequenced strain as a tree tip anchored temporally by its isolation date and connected by branches that trace its inferred evolutionary history. Generally, any mutations that occurred in a sequenced virus will be located at the tip of the tree. A mutation that occurs along an internal branch probably occurred once in a common ancestor of all strains that descend from that branch, rather than several independent times. For instance, the HA K160T mutation likely occurred once in a common ancestor of the 3c2.A viruses, sweeping through the H3N2 population to rise to fixation during the 2014-2015 season (Figure 2.1A).

However, when determining mutations that occurred in egg-passaged viruses, these assumptions are not necessarily true. It can only be assumed that an egg-passaged virus acquired just the mutations on the tree tip if it is sister to a NE-passaged strain (Figure

2.1B). If two egg-passaged viruses are sisters to each other and no NE-passaged virus, mutations that occur along the first internal branch likely actually occurred independently in both egg-passaged viruses (Figure 2.1C). This is because each egg-passaged strain derives from a different unpassaged virus rather than serial passaging in eggs. Thus, we assume that mutations occurring at internal branches with only egg-passaged descendants actually occur independently in each of those descendants.

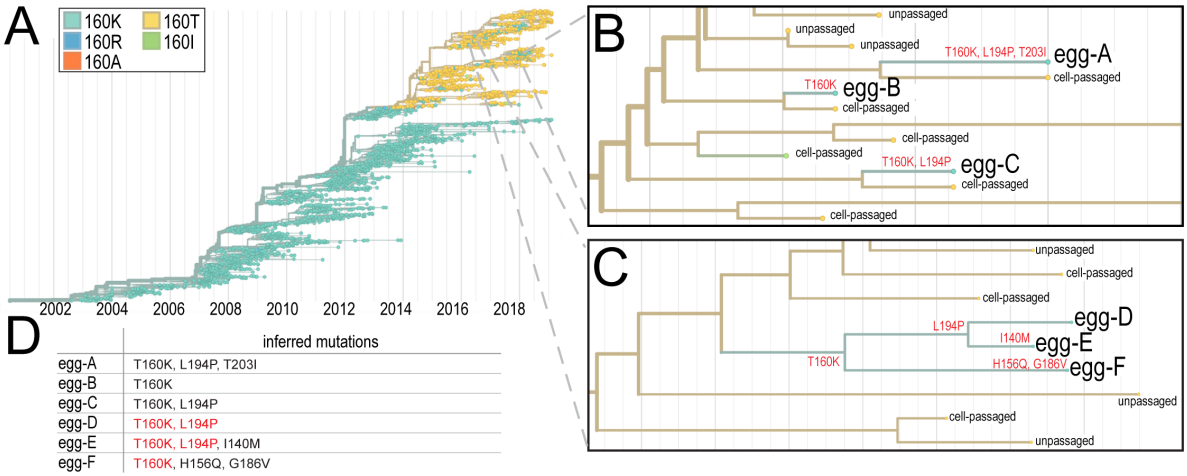

Figure 2.1: Egg-passaging mutations are inferred by a phylogeny-based method. A) H3N2 phylogeny, colored by HA 160 genotype. **B)** Egg-passaged mutations are equivalent to tip mutations for strains egg-A, egg-B, and egg-C, which are interspersed with non-egg-passaged strains. **C)** Egg-passaged viruses egg-D, egg-E, and egg-F cluster, placing some mutations that arise during egg-passaging on internal branches rather than at tree tips. The clustering method used in this paper groups these sequences into 5 clusters: egg-D, egg-E, egg-F, egg-D/egg-E, and egg-D/egg-E/egg-F. **D)** This method infers mutations for each egg-passaged sequence, including some that are on internal branches (red text).

Practically, we identified clusters of egg-passaged viruses and asserted that, during egg-passaging, each virus in the cluster accumulated the mutations occurring along the branch of the most recent common ancestor (MRCA) of the cluster. Clusters are defined as groups of 1 or more strains, meaning that this definition all includes tree tips. Additionally, clusters can be nested, allowing both mutations occurring along internal branches and at tree tips to be considered as occurring during egg-passaging (Figure 2.1C-D). See Methods for more details on this phylogenetic method to infer egg-passaging mutations.

The accuracy of this method can be assessed by comparing inferred mutations to known mutations. We took advantage of the 370 viruses that were sequenced as egg-passaged and

NE-passaged strains to determine "known" mutations by direct comparison of the pairs of sequences. Amongst these strains, we inferred 591 egg-passaging mutations in HA1, 21 of which were not real mutations. This is a false positive rate of 3.6%. Direct comparison of egg-passaged and NE-passaged viruses found that there 600 "known" egg-passaging mutations, 43 of which were not detected by our inference method (Figure 2.2A). This is a false negative rate of 7.2%.

Because this method uses phylogenetic relationships to infer egg-passaging mutations, errors in tree topology will result in errors in mutation inference. Hemagglutinin positions that have more entropy within the viral population (meaning a large portion of viruses have an alternate allele at this position) will tend to have a larger impact on tree topology. Conversely, residues that differ in only a couple viruses will have a small influence on topology. Because of this, we used our preliminary analysis to determine which positions mutate most commonly during egg-passaging and narrowed all subsequent analyses to consider only these sites. Notably, this lowered the number of false positives to 2/473, a rate of 0.4% (Figure 2.2A). Extrapolating the false positive rate to all egg-passaged viruses gives an expectation that, of the 617 inferred mutations, about 2.6 are false positives.

The two false positive mutations are a result of separation between the egg-passaged and NE-passaged strains on the phylogeny (Figure 2.2B-C). It is possible that common egg-passaging mutations can cause some egg-passaged viruses to appear more genetically similar to each other than to the unpassaged strains they were derived from. This would encourage egg-passaged viruses to cluster on the phylogeny, giving a false representation of shared ancestry. Despite skewing tree topology near the tips, egg clusters do not affect more ancestral nodes, meaning that the overall phylogenetic structure is not altered by the inclusion of such a large number of egg-passaged strains.

The low false positive rate and much higher false negative rate calibrate the method to infer mutations with a high level of confidence at the cost of missing some bona fide mutations. We opted for relatively conservative tuning to ensure that false positives do not

A

	All HA1 sites		Top egg mutation sites only	
	Between paired strains	All strains (estimated)	Between paired strains	All strains (estimated)
Inferred mutations	591	900	473	617
False positives	21	31.9	2	2.6
False positive rate	0.0355		0.0042	
"Known" mutations	600	924.3	496	764.1
False negatives	43	66.2	38	58.5
False negative rate	0.0717		0.0766	

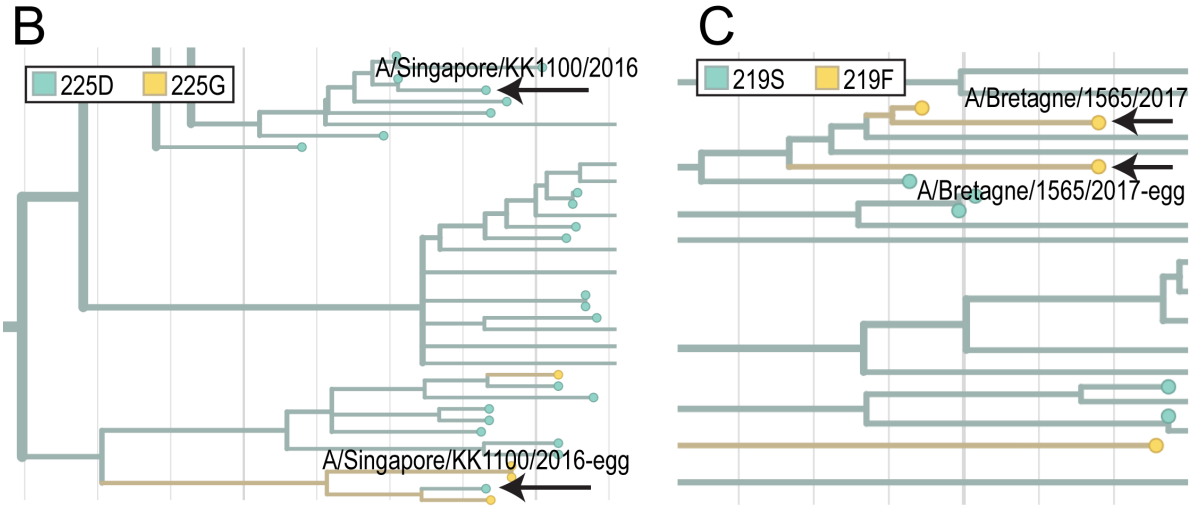

Figure 2.2: Inference of egg-passaging mutations is a conservative method with a low false positive rate. A) Evaluation of the false positive and negative rates of the method used in this paper to infer egg-passaging mutations. The false positive rate is very low when only the HA positions that are most commonly mutated in eggs are considered. These are sites 138, 156, 160, 186, 194, 203, 219, 225, and 246. Subsequent analyses are limited to only these sites, meaning only an estimated 2-3 inferred mutations are false positives. **B-C)** Snapshot of the topology surrounding the only two identified false positives.

skew future analyses and because frequently recurring mutations should be apparent across hundreds of viruses despite false negatives. Thus, this method confidently infers mutations in 570 egg-passaged viruses spanning about 16 years and a broad range of H3N2 diversity.

This phylogeny-based method enables mutations to be inferred in egg-passaged viruses that were not sequenced prior to egg-passaging and, additionally, provides clarity in determining mutations in egg-passaged viruses that can be directly compared to a NE-passaged strain. Only 116/370 of the egg-passaged strains paired with NE-passaged strains were sequenced from unpassaged virus (directly from patient samples). The remaining 254 pairs were sequenced after passaging in mammalian cells, making it difficult to determine whether sequence differences are due to mutations that occurred during cell passaging or during egg

passaging. A phylogeny-based method reduces this ambiguity by comparing the genotype of both sequences to other closely-related viruses.

2.2.2 Identify mutations specific to H3N2 viruses passaged in eggs

Simply identifying mutations that occurred during egg-passaging in individual viruses does not necessarily indicate that these mutations were a result of selection during egg-passaging, rather than just a result of genetic drift. Influenza A viruses have a high mutation rate of 2.0×10^{-6} mutations per nucleotide per infectious cycle (Nobusawa and Sato, 2006). During vaccine production, roughly 10^3-10^4 viral particles are passaged in eggs for about 7 cycles (Brauer and Chen, 2015). Given the length of HA (1701 nt), it can be expected that around 20-200 mutations arise in HA during vaccine production in eggs. Selectively-neutral mutations are subject to genetic drift, meaning they have a chance of disappearing and a chance of rising to high frequency in the viral population. However, mutations that confer a fitness advantage will selectively rise to high frequency in the viral population. Therefore, while individual strains may contain mutations that are due either to genetic drift or positive selection, a mutation that occurs in many strains can be expected to have a beneficial fitness effect. Moreover, a mutation that consistently occurs in egg-passaged viruses, but not cell-passaged or unpassaged viruses, can be expected to enhance viral fitness in chicken eggs.

Using mutations inferred by our phylogenetic method for 570 egg-passaged strains, we determined the most commonly-occurring egg-passaging mutations. All 11 of these mutations are significantly more frequent in egg-passaged versus cell- or unpassaged viruses (Figure 2.3A). The unique occurrence of these mutations in egg-passaged strains suggests that they are positively selected during egg-passaging, indicating that they are egg-adapted mutations. Throughout this paper the term "egg-adapted mutation" will be used to refer to these 11 mutations, while "egg-passaged mutation" signifies a substitution that likely occurred during egg-passaging due to drift. About 55% of egg-passaged viruses (315/570) have at least one of egg-adapted mutation.

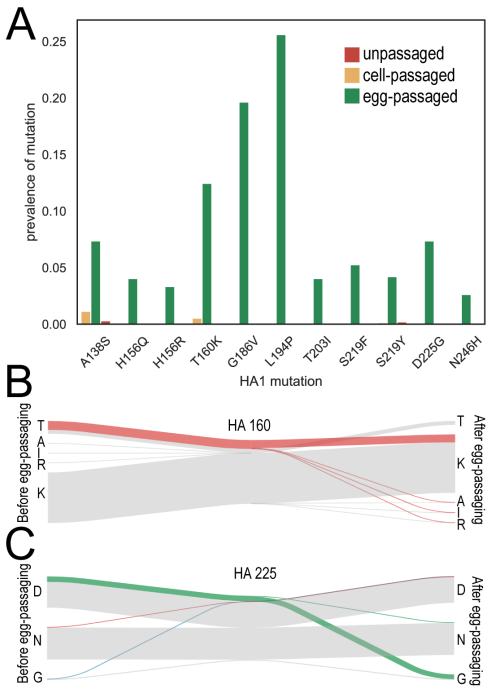

Figure 2.3: Eleven substitutions are true egg-adapted mutations. A) For each substitution, a chi-squared test confirms that the mutation is more prevalent among egg-passaged viruses versus cell-passaged or unpassaged viruses. **B-C)** Sankey plots show the genotypes of all 570 egg-passaged viruses before and after egg-passaging. Viruses that mutate are colored, and viruses that don't are shown in grey.

While most egg-adaptions mutate HA to a genotype not seen among circulating human H3N2 viruses, a few of these genotypes have been successful in human H3N2 as well. The T160K mutation, for instance, is considered a reversion mutation because prior to 2014 most human H3N2 viruses had a lysine at HA position 160. The 138S genotype is also found in a clade of human H3N2 viruses circulating from around 2013-present.

2.2.3 Clade-specificity of egg-adapted mutations

The T160K mutation provides a trivial example of how egg-adapted mutations can be limited to specific H3N2 strains. Prior to the emergence and fixation of HA 160T, all viruses had the HA 160K genotype, so egg-passaging could not have introduced a T160K mutation. It is well-documented that 160T viruses grow poorly in eggs and that the T160K reversion is a frequent egg-adaptation (Zost et al., 2017). This finding can be seen through analysis of the phylogeny as well: only viruses with threonine at site 160 mutate at this site during egg-

passaging (Figure 2.3B). This occurs despite an apparent tolerance for at least 5 different amino acids (T, K, A, I, R) at HA 160.

A similar, but slightly more interesting, case occurs at HA 225. Much of the H3N2 phylogeny has genotype HA 225N, where a D225G mutation would obviously be impossible. However, the N225G mutation is not found in any egg-passaged viruses, indicating that 225G is only selected for in certain contexts (Figure 2.3C). Additionally, these strains specifically acquire a D225G mutation (rather than a D225N reversion or some other mutation), indicating that 225G has a positive fitness effect rather than 225D having a negative effect. Together, this indicates that 225G is advantageous for the replication of specific strains in eggs.

To predict whether a given candidate vaccine virus is likely to mutate during egg-passaging, it is important to understand which egg-adapted mutations exhibit strain-specificity. To address this, we asked whether each egg-adapted mutation occurs preferentially on certain segments of the tree, or whether they are scattered more uniformly across the entire tree. Segments of the phylogeny are called clades (closely related viruses descending from a common ancestor). The WHO has defined clades of recent H3N2 viruses, however these clade definitions begin in 2011 and therefore fail to group nearly half of the strains on our phylogeny. Instead, we used mutations shared by large clusters of viruses to partition the full phylogeny into 18 clades (see Methods for more details). Each of these clades is defined by at least 1 amino acid substitution in HA and contain between 44-421 viruses (Figure 2.4A,C).

The overall amount of egg-adaptation increases over time (Figure 2.4B), reflecting increased reports of egg-passaging mutations in more recent H3N2 vaccine strains (Zost et al., 2017; Barr et al., 2018; Xue et al., 2016; Parker et al., 2016). This increase occurs specifically in certain strains, indicating that, in general, egg-adapted mutations occur preferentially in certain genetic backgrounds.

Clade-specificity of each mutation can be evaluated by an enrichment ratio, comparing the

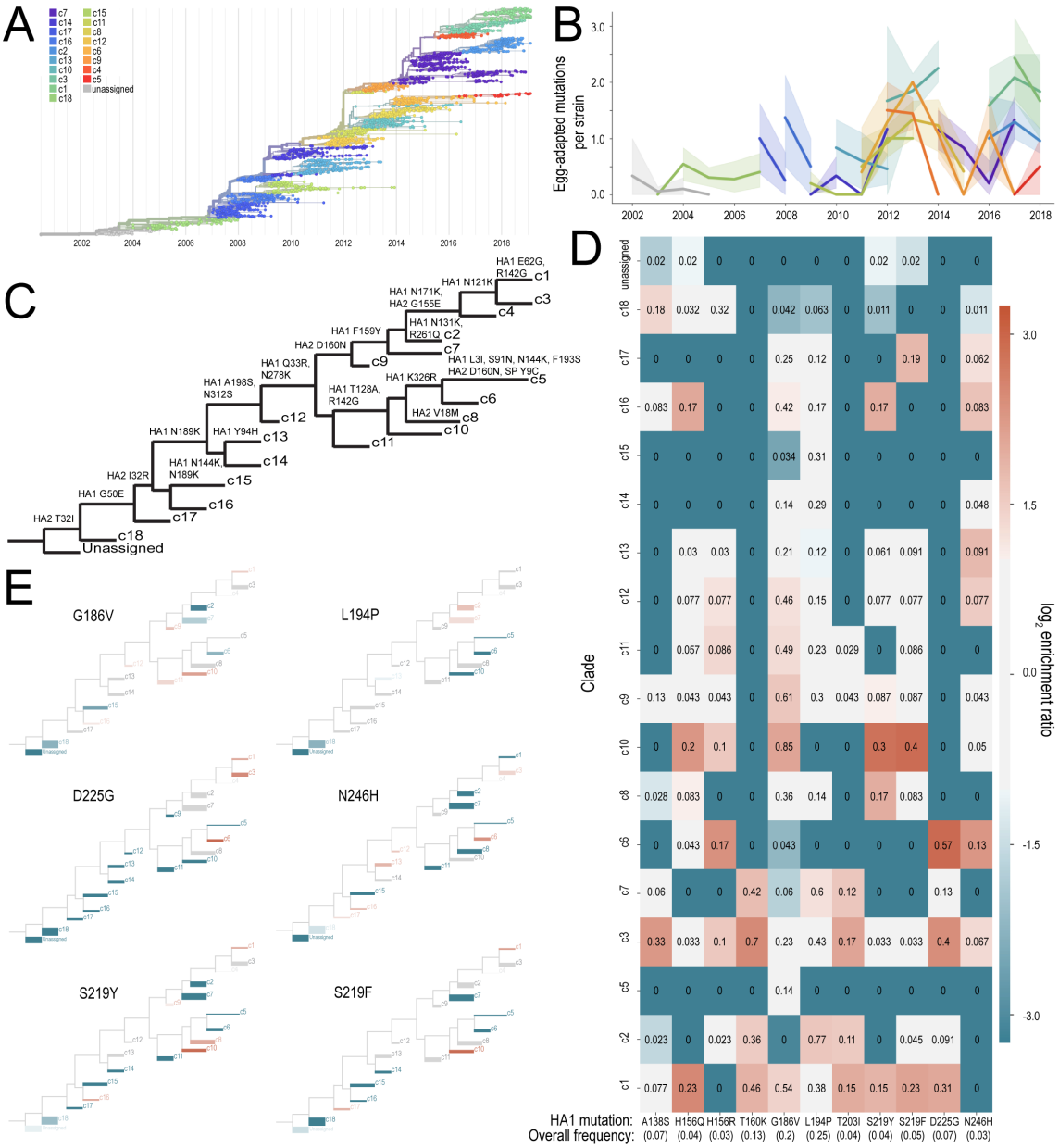

Figure 2.4: Egg-adapted mutations show some clade-specificity. A) Phylogeny colored by clade assignment. **B)** Egg-adaptation is more prevalent in more recent viruses and is clade-dependent. Egg-adaptation is quantified as the average number of bona fide egg-adapted mutations (the 11 mutations considered throughout this manuscript) per virus. Clades are colored according to the legend in (A) and shaded bands represent 95% confidence intervals. **C)** Cartoon showing unique mutations for each clade. **D)** A log enrichment ratio is computed for each mutation in each clade. Redder colors indicate the mutation is enriched in that clade and bluer colors indicate that it is depleted. The numbers in the middle of each box show the frequency of the corresponding mutation in that clade. The overall frequency of each mutation among all egg-passaged sequences is given in parentheses below the mutation on the x-axis. Clades containing fewer than 5 egg-passaged strains are excluded from this analysis. **E)** Cartoon clade maps are colored by log enrichment score (same scale as D). Branch thickness is proportional to the number of egg-passaged viruses in that clade.

frequency of the mutation in each clade to the frequency of the mutation in all egg-passaged viruses. A log enrichment ratio will be 0 if a mutation is present at the same frequency within a clade and throughout the entire tree. A positive log enrichment ratio indicates that the mutation is enriched in that clade, while a negative log enrichment ratio shows the opposite. Figure 2.4D displays log enrichment ratios as a heatmap, clearly revealing that egg-adapted mutations are not evenly distributed over the phylogeny.

It is apparent that certain clades of viruses, like c3, mutate quite often when they are passaged in eggs while others, like c5, very rarely do. Furthermore, G186V and L194P occur in viruses from nearly every clade, while N246H is confined to only half of the clades and D225G occurs in only 5 clades of viruses. Superimposing log enrichment ratios onto the tree topology makes these trends more apparent (Figure 2.4E). Figure 2.4E shows that, despite occurring in nearly every clade, L194P is most highly enriched in the closely-related clades c2 and c7 and is relatively rare in the closely-related clades c10 and c11, while G186V shows the opposite pattern.

Egg-adapted mutations at site 219 (S219F and S219Y) are most enriched in clades c1 and c10, both of which are defined by a R142G substitution. Clade c8, which descends from c10 and shares the 142G genotype, is also decently enriched for egg-adapted mutations at HA 219. However, no c5 or c6 viruses, which are also descendants of c10, mutate at position 219 during egg-passaging. This shows that while clade c1 and c10 viruses likely have similarities that encourage S219F/Y egg-adapted mutations, these similarities are not as simple as the single amino acid change that separates both of these clades from their ancestral clades. Strain-specificity may depend on the genotype at many HA residues and/or the genotype of the neuraminidase (NA) segment. This emphasizes how complicated strain-specificity of egg-adapted mutations likely is and argues for using viral clade to the predict the likelihood of mutation during egg-passaging.

2.2.4 Epistasis between egg-adapted mutations

Besides indicating clade-specificity of egg-adapted mutations, the log enrichment heatmap in Figure 2.4D also hints at relationships between egg-adapted mutations. For instance, S219F and H156Q appear to be highly enriched in the same clades. It is possible that epistasis between egg-adapted mutations also contributes to which HA sites mutate when a virus is passaged in eggs.

Epistasis, or the interaction between two HA residues, can have either positive or negative valence. In this case, a positive epistatic interaction could occur if the beneficial fitness effects of an egg-adapted mutation are enhanced by the co-occurrence of another egg-passaging mutation. Negative epistasis could indicate that the egg-adapted mutations are incompatible, either due to structural or functional constraints.

Within egg-passaged viruses, we considered pairwise epistatic interactions between the HA positions that commonly mutate during egg-passaging. In the absence of any epistatic interactions, the frequency of two genotypes co-occurring should be the product of the independent frequency of each genotype. Comparing this expected value to the actual observed frequency of co-occurrence gives the enrichment ratio (Figure 2.5A). This analysis points out several interactions, including positive epistasis between 225G and 138S, 194P and 203I, and 186V and 219F/Y. Negative epistatic interactions must be interpreted carefully from this analysis because, despite limiting the analysis to genotypes that occur in 5 or more strains, low enrichment scores can be an artifact of rare genotypes or can represent bona fide negative epistasis.

A striking example of negative epistasis occurs between 194P and 186V, the two predominant results of egg-adaptation. Zero H3N2 strains acquire both a L194P and a G186V mutation during egg-passaging (Figure 2.5C). The observation that 194P and 186V never co-occur could be a result of 1) direct negative epistasis between egg-adapted 194P and 186V, or 2) interactions between these two sites and the genetic background (i.e. the alleles that allow G186V do not allow L194P). There are 14 clades of egg-passaged viruses where

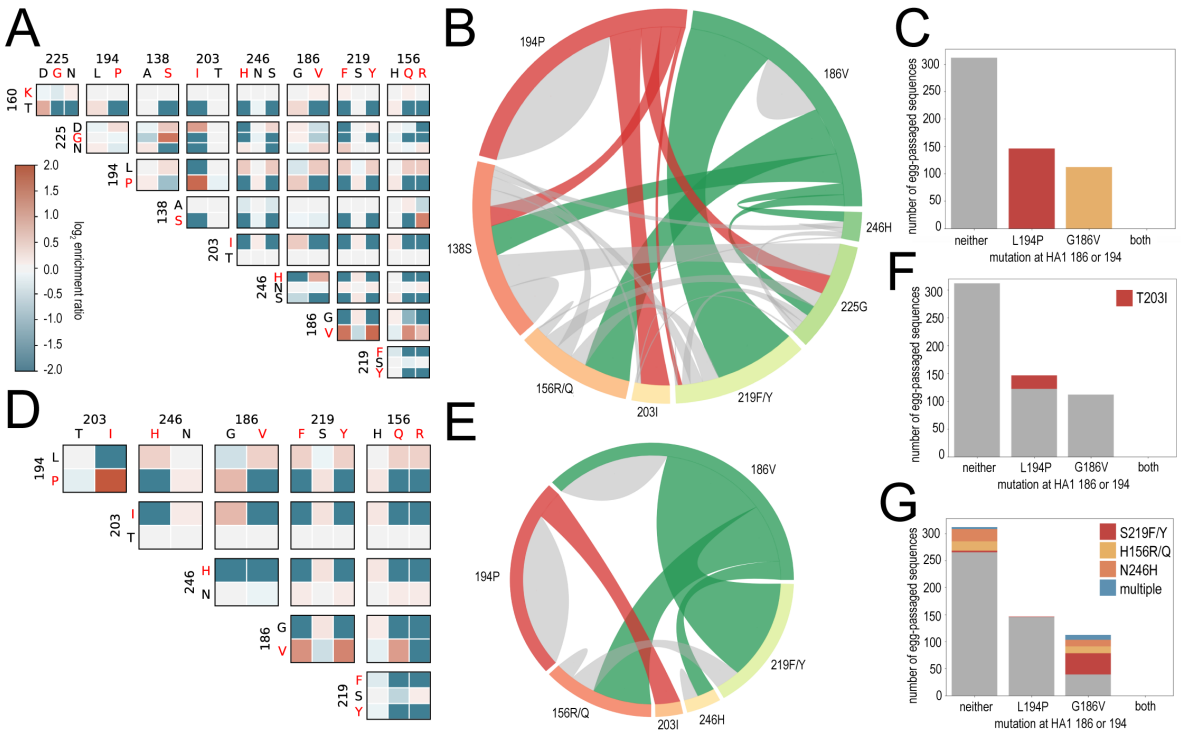

Figure 2.5: Epistatic interactions between egg-adapted mutations. A) Log enrichment ratio scores displayed as a heat map showing genotypes that co-occur in the same strain more commonly than expected in redder colors and genotypes that co-occur less often than expected in bluer colors. Only genotypes that appear in 5 or more strains are included. **B)** Chord diagram depicts the co-occurrence of two mutations. The L194P (red) and G186V (green) substitutions exhibit the strongest signs of epistasis with other mutations, so these interactions are emphasized by colored ribbons. Ribbon width is proportion to the number of strains that have both egg-adapted mutations. **C)** Despite their high prevalence among egg-passaged viruses, L194P and G186V mutations never co-occur. **D)** Epistatic enrichment ratio analysis is limited to only clades where both L194P and G186V occur at high rates (10% or higher). Enrichment for the co-occurrence of 194P/203I, 186V/219F, 186V/219Y, and 186V/156Q within these clades indicates bona fide epistasis. **E)** Epistatic interactions within the limited clades are shown by ribbons connecting genotypes. **F)** The T203I mutation (red) only occurs in egg-passaged strains that also have the L194P mutation. **G)** S219F/Y mutations (red) occur almost solely in egg-passaged strains that also have G186V and H156R/Q (yellow) and N246H (orange) never co-occur with L194P. Blue indicates sequences that have more than one of S219F/Y, H156R/Q, and N246H.

both L194P and G186V are observed, including 10 clades where both mutations are present in 10% or more of the viruses (Figure 2.4D). Because L194P and G186V never co-occur, despite being present in the same clades, this is probably due to direct negative epistasis. It is likely that genetic background can preferentially allow or prevent a G186V or L194P mutation during egg-passaging, and that once one of these mutations occur, direct negative epistasis prevents the other one.

During the preparation of this manuscript, Wu et al published experimental validation of

this negative epistasis between G186V and L194P (Wu et al., 2019). Their work reveals that co-occurance of G186V and L194P disrupts the HA receptor-binding site. This structural incompatibility prevents the virus from entering a cell and replicating.

Though G186V and L194P are mutually exclusive egg-adaptations, they both positively correlate with other egg-adapted mutations (Figure 2.5A). In fact, 67% of egg-passaged strains with the G186V mutation have at least 1 additional mutation (Figure 2.5B). Enrichment analysis shows a positive correlation between L194P and T203I and between G186V and H156R/Q, S219F/Y, and N246H. Again, these correlations could be a result of epistasis between egg-adapted mutations or a result of shared genetic background preference. To disambiguate between these possibilities, we limited our analysis of these potential interactions to clades where both G186V and L194P are present in at least 10% of egg-passaged viruses. Within these clades, enrichment ratios confirm that these correlations likely reflect epistatic interactions (Figure 2.5D-E).

It is possible that most of these pairwise correlations are mediated by epistatic interactions with HA site 194. All T203I mutations occur in viruses with a L194P mutation (Figure 2.5F), suggesting positive epistasis between 203I and 194P. Mutations at HA1 residues 156 and 246 occur in viruses with a G186V mutation and viruses with neither G186V nor L194P, but not viruses with a L194P mutation (Figure 2.5G). This is suggestive of negative epistasis between 194P and 156R/Q and 246H, though positive epistasis between these residues and 186V cannot be ruled out, and these two possibilities are not mutually exclusive. S219F/Y rarely occurs in strains that don't also have G186V, and especially high enrichment of S219F/Y with 186V indicates positive epistasis for these HA sites.

Other high log enrichment scores shown in Figure 2.5A did not hold up to similar secondary analyses of positive epistasis taking background specificity into account. So, no epistatic interactions were confidently identified between A138S, D225G, T160K and other egg-adapted mutations. The effects of these mutations is likely independent of other egg-adapted mutations: increased fitness during egg-passaging is not enhanced by the combina-

tion of these and other mutations (positive epistasis) nor do structural or functional viral constraints prevent them from co-occurring (negative epistasis).

2.2.5 Antigenic phenotype of egg-adapted mutations

Results of the epistasis analysis indicate that an H3N2 virus can undergo a couple potential mutational "pathways" to adapt to replication in eggs. These "pathways" are mainly defined by the L194P and G186V egg-adapted mutations. This observation is potentially encouraging that egg-adapted mutations may be able to be predicted and prevented by choosing a candidate vaccine virus that is incompatible with specific egg-adapted mutations. Egg-adapted mutations that should be prevented are those that result in antigenic changes that lower vaccine efficacy.

Mutation	Documented egg-passaged mutation	Attributed to VE decrease	HA domain	Viral replication in eggs	Antigenicity	References
L194P	Yes		190-helix, antigenic site B	Increase	Δ	(Wu et al., 2019, 2017; Chen et al., 2010; Levine et al., 2018)
G186V	Yes	Yes	Antigenic site B	Increase	Neutral	(Lu et al., 2005; Widjaja et al., 2006; Parker et al., 2016; Wu et al., 2019; Chen et al., 2010; Barman et al., 2015)
T160K	Yes	Yes	Antigenic site B, glycosylation	Increase	Δ	(Zost et al., 2017)
A138S	Yes		130-loop, antigenic site A			(Nakowitsch et al., 2014)
D225G	in H1N1		220-loop	Increase (H1N1)	Neutral (in H1N1)	(Xue et al., 2016)
S219F	Yes			Increase	Δ	(Widjaja et al., 2006; Parker et al., 2016)
S219Y	Yes	Yes		Increase	Neutral	(Parker et al., 2016)
T203I	Sequenced					(European Centre for Disease Prevention and Control, 2015; Worldwide Influenza Center: Francis Crick Institute, 2018)
H156Q	Yes	Yes	Antigenic site B	Increase	Δ	(Parker et al., 2016; Lin et al., 2010)
H156R	Yes		Antigenic site B	Increase		(Parker et al., 2016; Lin et al., 2010)
N246H	No					

Table 2.1: Documented phenotypic effects of egg-adapted mutations. Mutations that have not been described in any detail beyond appearing in published egg-passaged H3N2 sequences of egg-passaged viruses are denoted as "Sequenced". Egg-adapted mutations that were linked to low vaccine effectiveness are indicated. To our knowledge, there are no publications reporting D225G as an egg-adapted mutation in H3N2, though it has been documented to occur in H1N1. Where available, each mutation is marked as having an antigenic effect (Δ) or not (Neutral).

Several of the most common egg-adapted mutations we identified through phylogenetic methods have been documented by previous studies and some of this research has described the antigenic phenotype of these mutations as well (Table 2.1). The T160K substitution that is widely blamed for low effectiveness of the 2016-2017 vaccine is obviously reported to change antigenicity, as are egg-adapted mutations L194P, S219F and H156Q (Zost et al., 2017; Widjaja et al., 2006; Parker et al., 2016; Wu et al., 2017). Interestingly, G186V, which we identified as the second most predominant egg-adapted mutation is antigenically neutral.

However, the phenotypic effects listed in Table 2.1 are, again, only described for specific strains and it is not necessarily clear how many of these egg-adapted mutations affect

antigenicity more broadly, across all H3N2 strains. Again, we take a phylogenetic approach to infer antigenic phenotypes for all egg-passaged viruses. We employed a model that determines the relationship between genetic and antigenic evolution and uses this relationship to predict antigenicity across an entire phylogeny from known antigenic phenotypes (Neher et al., 2016). Antigenic differences between two viruses are measured by titers assays, which measure how well antibodies elicited against one virus neutralize the other virus. Titers are commonly measured by two different assays: hemagglutinin inhibition (HI) assay and focal reduction assay (FRA). Though HI titers have traditionally been used to compare H3N2 viruses, this assay does not work well with many recent strains for technical reasons. Because of this, we opted to use antigenic data from both the HI and FRA assays. Using titer measurements for tens of thousands of pairs of viruses and the phylogeny used throughout this manuscript, we ran the titers model to infer the antigenic phenotype of each egg-adapted mutation.

The model directly correlates amino acid substitutions with antigenic change and several egg-adapted mutations appear to alter H3N2 antigenicity (Table 2.2). In agreement with previous reports, G186V is antigenically neutral while L194P is not. Negative epistasis between 194P and 186V could provide an avenue to reduce the impact of egg-adapted mutations on VE: choosing a candidate vaccine virus with a valine at HA 186 could prevent L194P, thus, reducing the potential for antigenic change during egg-passaging. However, this strategy would not prevent other substitutions such as H156R, S219Y and T160K, which commonly co-occur with G186V and affect antigenicity.

To determine whether this strategy is likely to reduce potential antigenic change during egg-passaging, we compared antigenicity of viruses with the L194P mutation, to those with G186V or with neither substitution. This takes into account antigenic contributions from the genetic background and other egg-adapted substitutions. We used the titers model (Neher et al., 2016) to predict antigenic evolution from the root of the phylogeny to each virus located at the tips of the tree. The antigenic change from root to tip is measured in

Mutation	Antigenic units (2-fold decrease in titer) HI	FRA
L194P	0.3754	0.0289
G186V		
T160K		0.8675
A138S	0.0469	0.2602
D225G		0.8785
S219F		
S219Y		0.6530
T203I		
H156Q		
H156R	0.8369	0.4630
N246H		

Table 2.2: Predicted antigenic effect of each substitution.

antigenic units (1 unit = 2-fold decrease in titers), which takes into account the contribution of every substitution along this path. Therefore, antigenic change should increase with time as H3N2 diverges from the root, and because of this, it is not valid to compare strains sampled at different time points. However, the antigenic effects of egg-passaging can be evaluated through direct comparison of the 370 pairs of egg-passaged/NE-passaged viruses on this phylogeny.

For each pair of viruses, we calculated the difference in antigenic evolution between the viruses. Because these viruses have a shared evolutionary history until egg-passaging, the antigenic difference between paired strains should be attributable to egg-passaging. The effect of a specific egg-adapted mutation can be estimated by comparing egg-passaged viruses with just this substitution (and no other egg-adapted mutations) to their paired NE-passaged strains (Figure 2.6A). In agreement with previous reports and Table 2.2, L194P has a larger antigenic effect than G186V. The antigenic effect of G186V is not significantly different from the average effect of any single egg-passaged mutation.

As shown in Figure 2.5, these substitutions commonly co-occur with other egg-adapted mutations and, ultimately, VE is influenced by the antigenic change conferred by the combination of these co-occurring mutations. The combined antigenic effect of a specific substi-

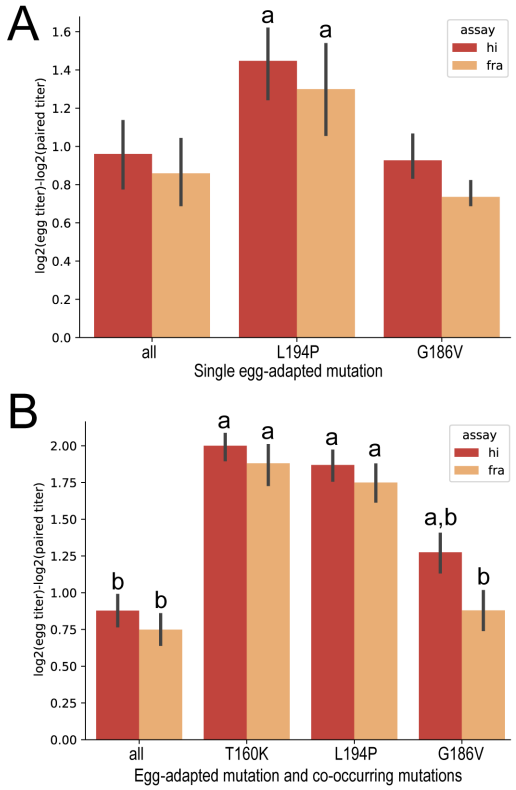

Figure 2.6: L194P mutational pathway results in larger antigenic change than the G186V pathway. A) The antigenic effect of single mutations is evaluated by computing the difference in predicted titer value between paired strains that differ by only one egg-passaging mutation. **B)** Predicted titers are compared between paired strains to estimate the antigenic effect of T160K, L194P, and G186V in combination with their co-occurring egg-adapted mutations. Statistical analysis is shown above each bar: t-test p-value ≤ 0.005 for viruses with the mutation compared to all egg-passaged sequences ("a"), or compared to viruses with T160K ("b"), is done separately for HI and FRA titer predictions.

tution and all co-occurring egg-adapted mutations can be evaluated by comparing all egg-passaged viruses that acquired this mutation (including viruses with multiple egg-adapted mutations) to their paired strains (Figure 2.6B). As expected, additional egg-adapted mutations increase antigenic distance between egg-passaged and NE-passaged viruses. The model predicts that, on average, egg-passaging increases antigenic change by 0.89 $\log_2$ HI titer units or 0.77 $\log_2$ FRA titer units. Viruses that did not mutate during egg-passaging differ from their paired strains by an average 0.007 antigenic units, which gives an estimation of the error of this method.

Egg-passaged strains with the G186V substitution are statistically more antigenically different than this overall average. Because G186V alone does not have an antigenic effect

(Table 2.1, Table 2.2, Figure 2.6A), co-occurring egg-adapted mutations must account for this antigenic change. Again, viruses with the L194P substitution are more antigenically divergent than those with G186V. This suggests that the egg-adaptation "pathway" involving L194P and T203I results in greater antigenic change than the "pathway" of G186V, H156R/Q, S219F/Y, and N246H mutations.

The most clear-cut example of an egg-adapted mutation that caused low VE through antigenic mismatch between the vaccine and circulating H3N2 viruses is the T160K mutation. Fittingly, the above phylogeny-based method predicts that the largest increase in antigenic distance between pairs of viruses occurs that acquire a T160K mutation during egg-passaging. The effects of other egg-passaged mutations can be benchmarked against the effect of T160K, which is known to alter VE. This suggests that egg-passaged viruses with a L194P mutation are more likely to decrease VE than viruses with a G186V substitution.

This relationship is especially evident when limiting this analysis to all strains that were sequenced before and after egg-passaging and which acquired only certain mutations during egg-passaging. This analysis reiterates the finding that H3N2 strains change more antigenically if they acquire only the L194P mutation during egg-passaging versus only the G186V mutation (Figure 2.7C). Additionally, mutations that are epistatically-associated with L194P or G186V further promote these antigenic differences, suggesting that the L194P "pathway" is more likely to decrease VE than the G186V "pathway".

2.3 Discussion

The influenza vaccine is predominantly manufactured in chicken eggs and, during this process, it is not uncommon for the vaccine virus to acquire mutations that increase fitness in eggs. Egg-adapted H3N2 mutations have been blamed for low VE during several flu seasons including, most notably, the 2016-2017 season (Zost et al., 2017). While egg-adaptation is a well-noted phenomena, reports of egg-adapted mutations are limited in scope to descriptions of specific mutations occurring in specific strains. In this manuscript, we employed phyloge-

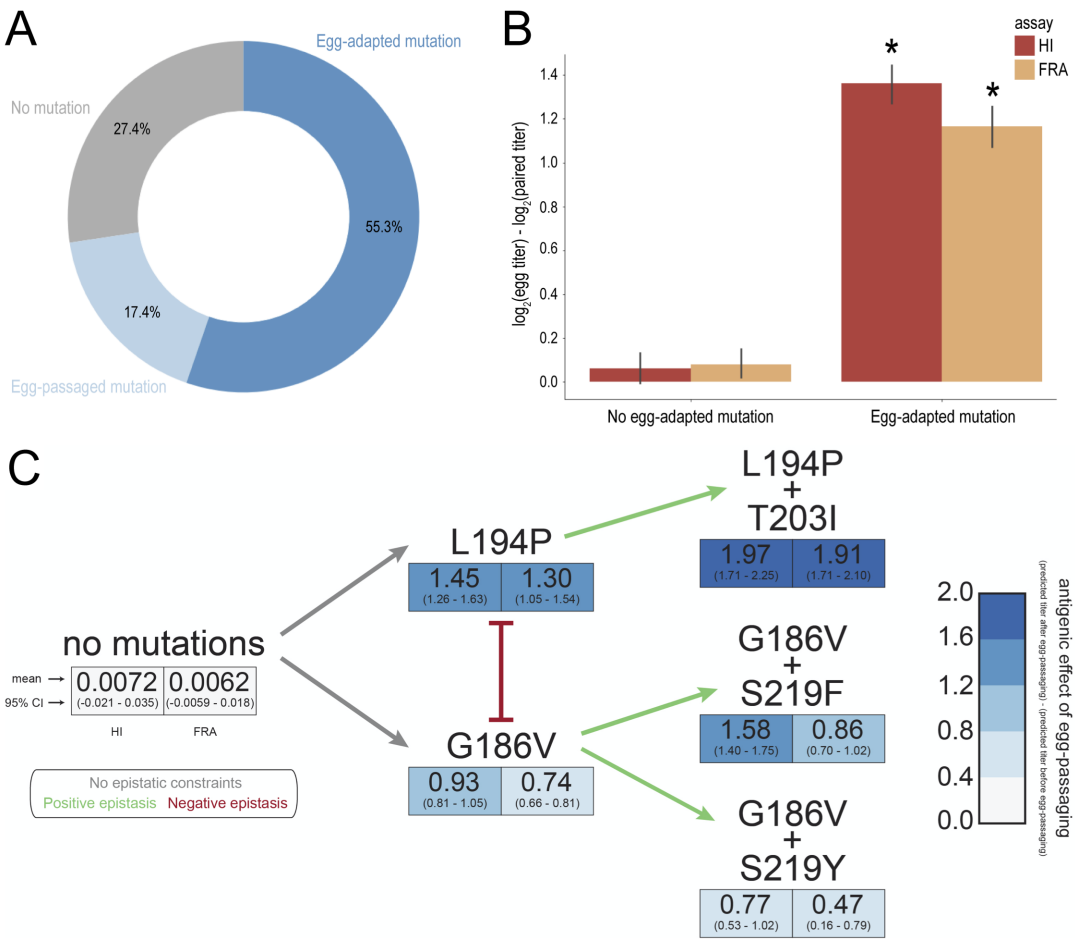

Figure 2.7: Egg-adapted mutations are common and alter antigenicity. **A)** Proportion of all egg-passaged viruses that acquired at least one egg-adapted mutation, an egg-passaged mutation but no egg-adapted mutation (distinction explained in Section 2.2), or no mutations at all during egg-passaged. **B)** Viruses with at least one egg-adapted mutation are predicted to be much more antigenically different than their NE-passaged paired strain, compared to viruses that do not have an egg-adapted mutation after egg-passaging. * indicates statistical difference from 0.0, which would be expected if a virus's antigenicity was not changed after egg-passaging (p-value > 0.001). **C)** Estimated antigenic impact of various epistatically-constrained pathways is computed from strains sequenced before and after egg-passaging that contain only the indicated mutations. Mean change in titers is estimated from HI (left) and FRA (right) assays and the magnitude of antigenic effect is color-coded with darker blue representing greater impact. Positive epistatic interactions are indicated by green arrows, negative epistatic interactions are shown by red bars, and gray arrows represent a lack of epistatic constraint.

netic methods to utilize the wealth of available genetic data, and consider broader patterns of H3N2 egg-adapted mutations and their effects on VE.

This gave an unprecedented perspective on just how prevalent egg-adaptation is in H3N2 viruses: about 73% of viruses mutate during egg-passaging and a conservative estimate indicates about 55% of strains acquire a bona fide egg-adapted mutation (Figure 2.7A). The

high frequency of these mutations is a critical concern for vaccine production because they have a striking impact on antigenicity. Egg-adapted mutations change the predicted titers of a virus by more than 2-fold (Figure 2.7B). In contrast, viruses that are passaged in eggs but do not develop an egg-adapted mutation are predicted to incur little-to-no antigenic change due to egg-passaging.

In addition to providing an overall view of the impact of egg-adaptation, this phylogenetic-based approach also reveals the prominence of some previously identified mutations, such as G186V, L194P and T160K (Figure 2.3A). While T160K is limited to more recent viruses, G186V and L194P are widespread over the H3N2 phylogeny (Figure 2.4D-E). The D225G mutation was found to be beneficial in the context of some genetic backgrounds but not others. Mutations at this position have also been noted to impact host-specificity in H1N1 (Wang et al., 2017; Yang et al., 2019). Additionally, considering such a large number of egg-passaged viral sequences, revealed previously undescribed egg-adapted mutations N246H and T203I. The N246H substitution appears in a variety of genetic backgrounds covering a large temporal range (Figure 2.4D-E), while T203I occurs only in viruses that also have a L194P egg-adapted mutation (Figure 2.5F).

Considering a large and diverse set of egg-passaged viruses exposes epistatic interactions, such as the positive epistasis between T203I and L194P, which may define disparate mutational "pathways" through which H3N2 adapts to replication in chicken eggs. The most striking of these interactions occurs between the 2 most prevalent and most topologically widespread mutations, G186V and L194P, which are completely mutually exclusive (Figure 2.5C). In addition to positive epistasis with T203I, the "L194P pathway" is also characterized by negative epistasis with H156R/Q, S219F/Y and N246H (Figure 2.5G). Besides L194P and T203I, the G186V substitution co-occurs with all other egg-adapted mutations, though the "G186V pathway" is specifically enriched in S219F/Y mutations, which rarely occur in the absence of G186V.

These distinct mutational "pathways" suggest that certain common egg-adapted muta-

tions can be selectively avoided by choosing candidate vaccine strains that occur preferentially on the other "pathway". Decisions to avoid certain substitutions while allowing others would be based on how the mutations affect viral effectiveness. Again, this manuscript builds on previous strain-specific studies of egg-adapted mutations to expose broader trends in their antigenic effects. This analysis indicates that the "L194P pathway" has a much larger antigenic effect than the "G186V pathway" (Figure 2.7C), and that it is similar in magnitude to the antigenic change caused by T160K (Figure 2.6B), which was responsible for the low VE during 2016-2017. While choosing a candidate vaccine virus that preferences the "G186V pathway" should result in an overall lower antigenic change, it is unclear exactly how this translates to the percent effectiveness of the vaccine. Also, the T160K substitution occurs on both pathways, so this strategy would not prevent the large antigenic impact of that egg-passaged mutation.

During the preparation of this manuscript, Wu et al published a report that also proposes mitigating potential antigenic change during egg-based vaccine production by preferencing 186V and, therefore, preventing the L194P mutation (Wu et al., 2019). Using X-ray crystallography, Wu and colleagues show that G186V and L194P are structurally incompatible HA mutations. They also confirm the differences in the antigenic effects of these two egg-adapted mutations with biolayer interferometry (BLI). This work provides strong experimental validation of conclusions we reached via phylogenetic methods, suggesting that our computational approach is an effective way to identify egg-adapted mutations, reveal epistatic interactions between them, and infer their antigenic effects.

The unique contribution of this research to the understanding of how egg-adapted mutations affect H3N2 VE is the consideration of a wide range of viruses, rather than a focus on specific strains. This methodology also has inherent drawbacks. Firstly, confidently identifying egg-adapted mutations relies on a large sample size and forcing the phylogeny to include so many egg-passaged viruses can skew the topology. Egg-passaged strains are directly derived from NE-passaged strains, meaning their closest relatives on the phylogeny should be

NE-passaged. However, not all egg-passaged strains were also sequenced as a NE-passaged virus, which can mean that the most genetically similar strain to a given egg-passaged virus is another egg-passaged virus. This clustering of egg-passaged viruses is amplified towards the tips of the tree, where strains are extremely genetically similar to each other and, thus, a given egg-adapted mutation that occurred separately in multiple strains will appear as an ancestral mutation during parsimonious phylogenetic reconstruction.

Because this tends to only skew the phylogeny at the tips of the tree, and does not affect the overall topology or major clade designations, this issue is dealt with by the phylogenetic method for inferring egg-passaging mutations explained in the "Phylogenetic method to infer egg-adapted mutations" section of the Results and Methods. However, this method would certainly be improved by the separation of egg clusters. This could be achieved either through the collection of more paired egg-passaged/NE-passaged sequences or through masking specific HA residues that mutate during egg-passaging when the phylogenetic tree is built. Masking would prevent these sites from influencing the tree topology, but may not be useful because many egg-adapted mutations occur at HA positions (like 138, 160, 225) that have mutated elsewhere in the evolutionary history of circulating H3N2.

Secondly, this phylogeny-based method of egg-passaged mutations uses clades to determine the genetic background specificity of each egg-adapted mutation. These correlations are clearly dependent on how clades are defined, so conclusions about clade specificity should be understood as an enrichment or depletion of a specific mutation in viruses located in those general regions of the phylogeny, and not necessarily as viruses that precisely fit into the clades defined by this study.

Third, the phylogenetic analysis in this manuscript is based solely on the sequence of H3N2 HA. This decision was made because the sequence of HA is available for almost all strains while the full genome sequence is not, and because HA is the primary location of ongoing viral evolution and, specifically, of egg-adapted mutations. However, where possible, using the entire H3N2 genome could help separate egg-passaged clusters on the phylogeny

and provide a more comprehensive analysis of background specificity and of epistasis.

2.4 Methods

All code, data, and results for this project can be found in the following Github repository: `https://github.com/blab/egg-passage`

2.4.1 Sequence and titers data

Publicly accessible influenza H3N2 sequence and titers data were downloaded from Nextstrain's Fauna database (Hadfield et al., 2018). This data was originally collected from sources such as NCBI (`https://www.ncbi.nlm.nih.gov/`), GISAID (`https://gisaid.org`, (Shu and McCauley, 2017)) and ViPR (`https://www.viprbrc.org`). From these databases, a subset of sequences was selected to build a phylogeny and perform subsequent analyses. To facilitate egg-passaging analyses, this subset was enriched for egg-passaged sequences sampled between 2002-2019 (n=570), and all available non-egg-passaged sequences that match these egg-passaged strains (n=370). These strains were identified by `find_egg_seqs.py`. The remaining sequences were selected by Nextstrain's subsampling algorithm, which ensures even temporal and geographic sampling. Titers data from "cell" and "egg" databases was concatenated into a database of 2969 HI titer measurements and another of 1222 FRA measurements.

2.4.2 Phylogenetic tree

The time-resolved phylogenetic tree was constructed with Nextstrain's Augur (Huddleston et al., 2021), IQ-TREE 2.1.2 (Nguyen et al., 2015), and TreeTime 0.8.2 (Sagulenko et al., 2018). Augur was run twice, once with HI titer data and once with FRA data, using 12 year resolution. The resulting phylogeny was visualized using Nextstrain Auspice (Hadfield et al., 2018). All subsequent analyses use the `.json`, `.fasta`, `.tsv`, and `.nwk` files output by Augur.

2.4.3 Phylogenetic method to infer egg-adapted mutations

A Python script, `organize_output.py`, was written to extract relevant information (such as strain name, isolation date, titer predictions, and mutations on tree tips) from Augur output files and organize it into Pandas dataframes. Using a traversal function, `organize_output.py` keeps track of each virus's mutational history from tree root to tip.

Starting at the tree root and walking forward in time, groups are formed from viruses that share the same mutational history up until that point. If, at any point, all viruses in a group are egg-passaged, the mutations occurring on the most recent branch are considered to be egg-passaging mutations that occurred separately in each virus in the group. This process is done iteratively, so groups of egg-passaged viruses can be nested. All mutations occurring at tree tips are considered groups of size 1 and, therefore, are considered egg-passaging mutations.

If the only non-egg-passaged strains in a group subsequently revert the group-defining mutation, this group-defining mutation is also considered to have occurred separately in each egg-passaging strain in the group. This prevents false negative mutation calls caused by inaccurate topology, which probably results from phylogenetic construction based on an assumption of parsimony. Because each egg-passaged sequence derives from a non-passaged sequence (rather than serial passaging in eggs), it is actually more likely that each of these strains acquired the mutation separately, rather than the non-egg-passaged sequence mutating and then acquiring a reversion.

If grouping results in an egg-passaged sequence that has two mutations at the same HA position, the egg-passaging mutation at that position is inferred to be from the original amino acid in the first mutation to the mutated amino acid in the last mutation. This can include reversion mutations. For example, an egg-passaged strain that has I140R and R140K mutations according to the grouping protocol, is inferred to actually have an I140K mutation. Similarly, an egg-passaged strain with both D225G and G225D, would be inferred to have no mutation at position 225. This prevents false positive mutation calls that result

from an inaccurate topology, skewed by the large number of egg-passaged sequences.

2.4.4 Validation of phylogenetic mutation inference method

The validity of this phylogenetic method to infer egg-passaged mutations was assessed by calculating the false positive and negative rates (Figure 2.2A). False positives were identified through direct comparison of viruses that were sequenced both as an egg-passaged strain and as either an unpassaged or cell-passaged strain. Every inferred egg-passaging mutation was, thus, validated by comparing the sequence of the egg-passaged virus to its paired strain. Inferred mutations that were not confirmed by this method were considered false positives. False positives were separately evaluated at all HA sites and at only the sites that mutate most commonly during egg-passaging. The latter greatly reduces the false positive rate.

False negatives were evaluated by determining every difference between each pair of sequences. If the difference corresponded to a mutation occurring at the tree tip of the unpassaged or cell-passaged strain, it was considered to be a mutation in that strain. All other differences were considered to be "known" egg-passaging mutations. Every "known" mutation that was not inferred was considered a false negative. The false negative rate was also calculated separately at all HA sites and at only the sites that mutate most commonly during egg-passaging, and the rate was consistent between these two.

2.4.5 Identification of mutations specific to egg-passaging

The most common egg-passaging mutations inferred in all 570 egg-passaged strains were determined by `find_egg_mutations.py`. The prevalence of these mutations in egg-passaged viruses was compared to cell-passaged and unpassaged viruses is compared using a chi-squared test.

2.4.6 Clade designations

Closely-related viruses were grouped into clades based on the mutational history of each virus (described above) by `assign_clades.py`. At each branch in the path from tip to root, groups of viruses with identical mutational history were formed. If the group met any of the following criteria (evaluated in this order), it was considered a clade:

1. 20 or more members, and 3 or more clade-defining amino acid mutations
2. 50 or more members, and 2 or more clade-defining amino acid mutations
3. 100 or more members, and at least 1 clade-defining amino acid substitution

Groups were iteratively evaluated on these criteria at each point walking backward in time, so that more recent strains will form clades that descend from older sequences. Strains near the root that are not included in any of these clades are put in the 'unassigned' clade.

2.4.7 Determination of clade specificity of mutations by enrichment ratios

Enrichment or depletion of egg-adapted mutations in each clade was assessed and plotted as a heatmap by `plot_background.py`. The enrichment of mutation A in clade 1 is:

$$\log_2 \frac{f_{A1}}{f_A}$$

where f_{A1} is the frequency of mutation A in clade 1 and f_A is the frequency of mutation A in all egg-passaged sequences.

2.4.8 Identification of epistatic interactions between egg-adapted mutations

Pairwise epistatic interactions are evaluated by the observation that two egg-adapted mutations occur in the same virus more or less commonly that would be expected. This is evaluated for mutations A and B through the following enrichment ratio:

$$\log_2 \frac{f_{AB}}{f_A * f_B}$$

where f_{AB} is the frequency of viruses that have both mutations, f_A is the overall frequency of viruses with mutation A, and f_B is the overall frequency of viruses with mutation B. These values are plotted as a heatmap by `plot_egg_epistasis.py`. This script also plots chord diagrams to highlight specific interactions with G186V and L194P.

2.4.9 Prediction of antigenic change attributed to egg-adapted mutations

Augur (Huddleston et al., 2021) was used to implement the substitution model to predict titers for each virus on the phylogeny (Neher et al., 2016). This was done both using FRA titers data and HI titers data. The antigenic effect of each egg-adapted mutation was determined from the Augur output file, which lists the effect attributed to each substitution by the model. The antigenic change from the tree root ('cTiterSub') was also extracted from Augur output files compared between paired strains by `plot_titer_diffs.py`. Differences were tested for statistical significance using a student's t-test.

Chapter 3

Evidence for adaptive evolution in seasonal coronaviruses

This work was originally published with *eLife* at https://doi.org/10.7554/eLife.64509.

3.1 Introduction

Coronaviruses were first identified in the 1960s and, in the decades that followed, human coronaviruses (HCoVs) received a considerable amount of attention in the field of infectious disease research. At this time, two species of HCoV, OC43 and 229E, were identified as the causative agents of roughly 15% of common colds (McIntosh, 1974; Heikkinen and Järvinen, 2003). Infections with these viruses were shown to exhibit seasonal patterns, peaking in January-March in the Northern Hemisphere, as well as yearly variation, with the greatest incidence occurring every 2-4 years (Monto and Lim, 1974; Hamre and Beem, 1972). Subsequently, two additional seasonal HCoVs, HKU1 and NL63, have entered the human population. These 4 HCoVs endemic to the human population usually cause mild respiratory infections, but occasionally result in more severe disease in immunocompromised patients or the elderly (Liu et al., 2020a). In the past 20 years, three additional HCoVs (SARS-CoV-1, MERS-CoV and SARS-CoV-2) have emerged, which cause more severe respiratory illness. At the writing of this paper, amidst the SARS-CoV-2 pandemic, no vaccine for any HCoV is currently available, though many candidate SARS-CoV-2 vaccines are in production and clinical trials (Krammer, 2020).

Coronaviruses are named for the ray-like projections of spike protein that decorate their surface. Inside these virions is a positive-sense RNA genome of roughly 30kB (Li, 2016). This large genome size can accommodate more genetic variation than a smaller genome

(Woo et al., 2009). Genome flexibility, coupled with a RNA virus error-prone polymerase (Drake, 1993) and a high rate of homologous recombination (Pasternak et al., 2006), creates genetic diversity that is acted upon by evolutionary pressures that select for viral replication. This spawns much of the diversity within and between coronaviruses species (Woo et al., 2009; Hon et al., 2008), and can contribute to the virus' ability to jump species-barriers, allowing a previously zoonotic CoV to infect and replicate in humans.

The battle between virus and host results in selective pressure for mutations that alter viral antigens in a way that evades immune recognition. Antigenic evolution, or antigenic drift, leaves a characteristic mark of positively selected epitopes within the viral proteins most exposed to the host immune system (Smith et al., 2004). For CoVs, this is the spike protein, exposed on the surface of the virion to human humoral immunity. Some human respiratory illnesses caused by RNA viruses, like seasonal influenza (Smith et al., 2004), evolve antigenically while others, like measles, do not (Fulton et al., 2015). Because of this, seasonal influenza vaccines must be reformulated on a nearly annual basis, while measles vaccines typically provide lifelong protection. Whether HCoVs undergo antigenic drift is relevant not only to understanding HCoV evolution and natural immunity against HCoVs, but also to predicting the duration of a vaccine's effectiveness.

Early evidence that closely-related HCoVs are antigenically diverse comes from a 1980s human challenge study in which subjects were infected and then reinfected with a variety of 229E-related strains (Reed, 1984). All subjects developed symptoms and shed virus upon initial virus inoculation. After about a year, subjects who were re-inoculated with the same strain did not show symptoms or shed virus. However, the majority of subjects who were re-inoculated with a heterologous strain developed symptoms and shed virus. This suggests that immunity mounted against 229E viruses provides protection against some, but not all, other 229E strains. This is a result that would be expected of an antigenically evolving virus.

More recent studies have identified 8 OC43 genotypes and, in East Asian populations, certain genotypes were shown to temporally replace other genotypes (Lau et al., 2011; Zhang

et al., 2015; Zhu et al., 2018). Whether certain genotypes predominate due to antigenic differences that confer a fitness advantage is not known. However, evidence for selection in the spike protein of one of these dominant OC43 genotypes has been provided by d_N/d_S, a standard computational method for detecting positive selection (Ren et al., 2015). This method has also been used to suggest positive selection in the spike protein of 229E (Chibo and Birch, 2006). Additionally, two genetically distinct groupings (each of which include multiple of the aforementioned 8 genotypes) of OC43 viruses have been shown to alternate in prevalence within a Japanese community, meaning that the majority of OC43 infections are caused by one group for about 2-4 years at which point the other group begins to account for the bulk of infections. It has been suggested that antigenic differences between these groups contribute to this epidemic switching (Komabayashi et al., 2020).

However, a similar surveillance of the NL63 genotypes circulating in Kilifi, Kenya found that NL63 genotypes persist for relatively long periods of time, that people become reinfected by the same genotype, and that reinfections are often enhanced by prior infection (Kiyuka et al., 2018). These findings are inconsistent with antigenic evolution in NL63.

Here, we use a variety of computational approaches to detect adaptive evolution in spike and comparator proteins in HCoVs. These methods were designed as improvements to d_N/d_S with the intention of identifying adaptive substitutions within a serially-sampled RNA virus population. We focus on the seasonal HCoVs that have been continually circulating in humans: OC43, 229E, HKU1 and NL63. Our analyses of nonsynonymous divergence, rate of adaptive substitutions, and Time to Most Recent Ancestor (TMRCA) provide evidence that the spike protein of OC43 and 229E is under positive selection. Though we conduct these analyses on HKU1 and NL63, we do not observe evidence for adaptive evolution in the spike protein of these viruses. For HKU1, there is not enough longitudinal sequencing data available for us to confidently make conclusions as to whether or not this lack of evidence reflects an actual lack of adaptive evolution.

3.2 Results

3.2.1 Phylogenetic consideration of viral diversity and recombination in OC43 and 229E

We constructed time-resolved phylogenies of the OC43 and 229E using publicly accessible sequenced isolates. A cursory look at these trees confirms previous reports that substantial diversity exists within each viral species (Zhang et al., 2015; Komabayashi et al., 2020; Lau et al., 2011). Additionally, the trees form ladder-like topologies with isolate tips arranged into temporal clusters rather than geographic clusters, indicating a single global population rather than geographically-isolated populations of virus. The phylogeny of OC43 bifurcates immediately from the root (Figure 3.1), indicating that OC43 consists of multiple, co-evolving lineages. Because of the distinct evolutionary histories, it is appropriate to conduct phylogenetic analyses separately for each lineage. We have arbitrarily labeled these lineages 'A' and 'B' (Figure 3.1).

Because recombination is common amongst coronaviruses (Pasternak et al., 2006; Hon et al., 2008; Lau et al., 2011), we built separate phylogenies for each viral gene. In the absence of recombination, each tree should show the same evolutionary relationships between viral isolates. A dramatic difference in a given isolate's position on one tree versus another is strongly indicative of recombination (Kosakovsky Pond et al., 2006). Comparing the RNA-dependent RNA polymerase (RdRp) and spike trees reveals this pattern of recombination in some isolates (Figure 3.S1A). A comparison of the trees of the S1 and S2 sub-domains of spike shows more limited evidence for intragenic recombination (Figure 3.S1B), which is consistent with the fact that the distance between two genetic loci is inversely-related to the chance that these loci remain linked during a recombination event. Though intragenic recombination likely does occur occasionally, analyzing genes, rather than isolates, greatly reduces the contribution of recombination to genetic variation in our analyses.

Thus, in all of our analyses, we use alignments and phylogenies of sequences of single genes (or genomic regions) rather than whole genome sequences of isolates. We designate

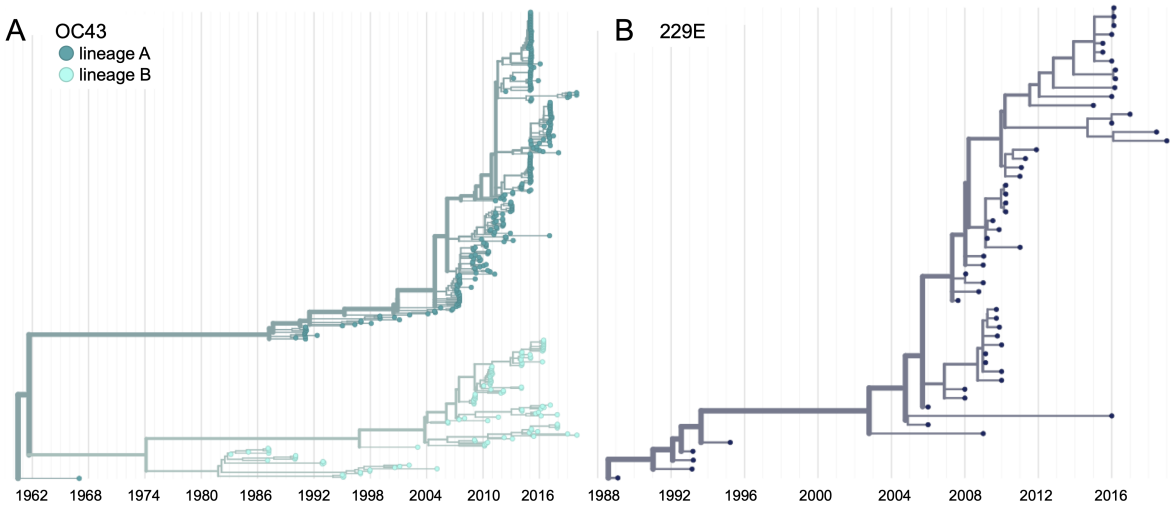

Figure 3.1: Phylogenetic trees for spike gene of seasonal HCoVs OC43 and 229E. Phylogenies built from **A)** OC43 spike sequences from 389 isolates over 53 years, and **B)** 229E spike sequences from 54 isolates over 31 years. OC43 bifurcates immediately after the root and is split into two lineages: lineage A (dark teal) and lineage B (light teal). 229E and contains just one lineage (dark blue). For the analyses in this paper, the evolution of each gene (or genomic region) is considered separately, so phylogenies are built for each viral gene, and those phylogenies are used to split isolates into lineages for each gene. These are temporally resolved phylogenies with year shown on the x-axis. The clock rate estimate is 5×10^{-4} substitutions per nucleotide site per year for OC43 and 6×10^{-4} for 229E.

the lineage of those genes (or genomic regions) based on the gene's phylogeny. Though most isolates contain all genes from the same lineage, some isolates have, say, a lineage A spike gene and a lineage B RdRp gene. This strategy allows us to consider the evolution of each gene separately, and interrogate the selective pressures acting on them.

It is worth noting that the analyses we use here to detect adaptive evolution canonically presume that selective pressures are acting on single nucleotide polymorphisms (SNPs). However, it is possible that recombination also contributes to the genetic variation that is acted on by immune selection. This would be most likely to occur if two closely-related genomes recombine, resulting in the introduction of a small amount of genetic diversity without disrupting crucial functions. Our analyses do not aim to determine the source of genetic variation (i.e. SNPs or recombination), but rather focus on identifying if and how selection acts on this variation.

Because of its essential role in viral replication and lack of antibody exposure, we expect RdRp to be under purifying selection to maintain its structure and function. If HCoVs

evolve antigenically, we expect to see adaptive evolution in spike, and particularly in the S1 domain of spike (Hofmann et al., 2006; Hulswit et al., 2019), due to its exposed location at the virion's surface and interaction with the host receptor. Mutations that escape from population immunity are beneficial to the virus and so are driven to fixation by positive selection. This results in adaptive evolution of the virus population.

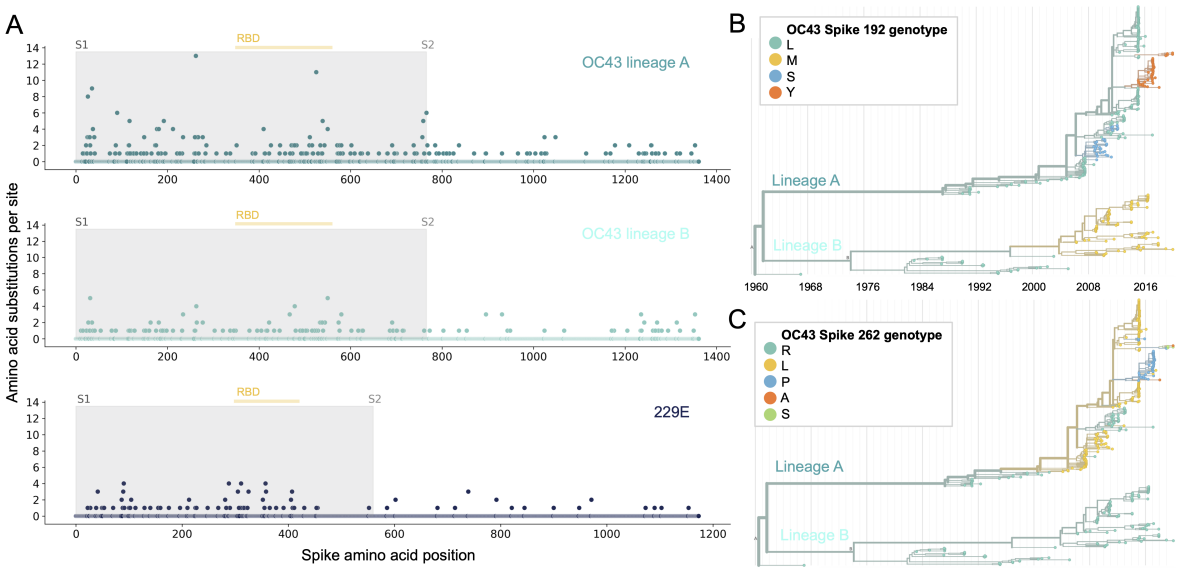

Figure 3.2: More sites mutate repeatedly within spike S1 versus S2. A) Number of substitutions observed at each amino acid position in the spike gene throughout the phylogeny. S1 (gray) and S2 (white) are indicated by shading and the number of substitutions per site is indicated by a dot and color-coded by HCoV lineage. The putative receptor-binding domains for 229E (Li et al., 2019) and the putative domain for OC43 (Lau et al., 2011) are indicated with light yellow bars. Asterisks indicate two example positions (192 and 262), which mutate repeatedly throughout the OC43 lineage A phylogeny. The OC43 phylogeny built from spike sequences and color-coded by genotype at position 192 and 262 is shown in **B)** and **C)**, respectively.

3.2.2 Phylogenetic inference of substitution prevalence within spike

Using phylogenies constructed from the spike gene, we tallied the number of independent amino acid substitutions at each position within spike. The average number of substitutions per site is higher in S1 than S2 for HCoV lineages in OC43 and 229E (Figure 3.2A). We focus on S1 rather than the Receptor-Binding Domain (RBD) within S1 in our analyses, because it is known that neutralizing antibodies bind to epitopes within the N-Terminal Domain (NTD) as well as the RBD of S1 (Liu et al., 2020b; Zhang et al., 2018; Zhou et al.,

2019). A greater occurrence of repeated substitutions is expected if some mutations within S1 confer immune avoidance. Alternatively, these repeated substitutions could be a result of high mutation rate and random genetic drift as has been shown at particular types of sites in SARS-CoV-2 (van Dorp et al., 2020). However, this latter hypothesis should affect all regions of the genome equally and should not result in a greater number of repeated substitutions in S1 than S2.

If the repeated mutations are a product of immune selection, not only should S1 contain more repeated mutations, but we would also expect these mutations to spread widely after they occur due to their selective advantage. Additionally, we expect sites within S1 to experience diversifying selection due to the ongoing arms race between virus and host immune system. This is visible in the distribution of genotypes at the most repeatedly-mutated sites in OC43 lineage A (Figure 3.2B and 3.2C).

3.2.3 Nonsynonymous and synonymous divergence in RdRp and subdomains of spike

An adaptively evolving gene, or region of the genome, should exhibit a high rate of nonsynonymous substitutions. For each seasonal HCoV lineage, we calculated nonsynonymous and synonymous divergence as the average Hamming distance from that lineage's most recent common ancestor (Zanini et al., 2015). The rate of nonsynonymous divergence is markedly higher within spike versus RdRp of 229E and OC43 lineage A (Figure 3.3A). While nonsynonymous divergence increases steadily over time in spike, it remains roughly constant at 0.0 in RdRp. These results suggest that there is predominantly positive selection on OC43 and 229E spike, but predominantly purifying selection on RdRp. Separating spike into the S1 (receptor-binding) and S2 (membrane-fusion) domains reveals that the majority of nonsynonymous divergence in spike occurs within S1 (Figure 3.3B). In fact, the rates of nonsynonymous divergence in S2 are similar to those seen in RdRp, suggesting S2 evolves under purifying selection while S1 evolves adaptively.

Though we would expect synonymous divergence to be equivalent in all areas of the

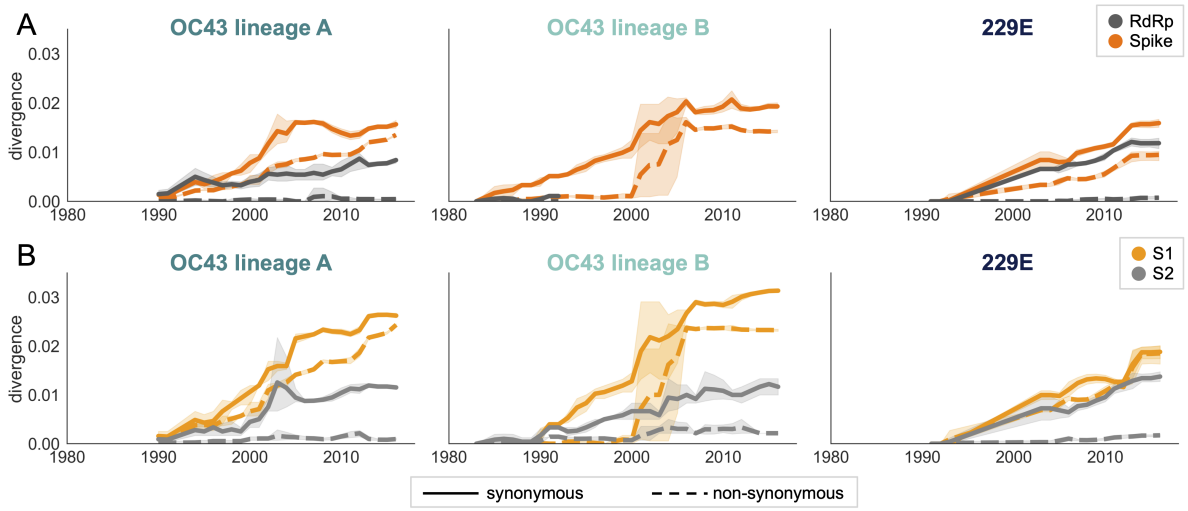

Figure 3.3: Nonsynonymous divergence is higher in OC43 and 229E Spike S1 versus S2 or RdRp. A) Nonsynonymous (dashed lines) and synonymous divergence (solid lines) of the spike (dark orange) and RdRp (dark gray) genes of all 229E and OC43 lineages over time. Divergence is the average Hamming distance from the ancestral sequence, computed in sliding 3-year windows which contain at least 2 sequenced isolates. Shaded region shows 95% confidence intervals. Note that the absence of a line means there fewer than 2 sequences available at this timepoint and that, therefore, the divergence is not calculated. **B)** Nonsynonymous and synonymous divergence within the S1 (light orange) and S2 (light gray) domains of spike. Year is shown on the x-axis and is shared between plots.

genome, this is not born out in our results. It is unclear whether the difference in synonymous divergence between genes reflects an actual biological difference. However, the ratio of nonsynonymous divergence in spike to nonsynonymous divergence in RdRp is consistently higher than the equivalent ratio of synonymous divergence (Figure 3.S5). Thus, despite differences in synonymous divergence, spike is accumulating more relatively more nonsynonymous divergence than RdRp.

We compared our analysis of divergence to the results a more standard approach for detecting positive selection on certain branches of a phylogeny. This approach, called MEME, is maximum-likelihood method which gives a single d_N/d_S value for each gene (Murrell et al., 2012; Weaver et al., 2018). In agreement with measures of nonsynonymous divergence over time, d_N/d_S estimates are higher in Spike than RdRp and higher in S1 than S2 (Table 3.1). Our estimate of d_N/d_S in OC43 spike is similar to the previously reported estimate of roughly 0.3 (Ren et al., 2015). However, we believe the standard d_N/d_S approach is not the ideal tool for detecting adaptive evolution in HCoVs because it is a phylogenetic approach, which

may be biased by recombination, and also because some assumptions of the model hold true for mammalian genomes, but not necessarily for RNA viruses

	RdRp	Spike	S1	S2
229E	0.143	0.441	0.662	0.166
OC43 lineage A	0.080	0.435	0.466	0.301
OC43 lineage B	0.061	0.317	0.418	0.234
NL63	0.068	0.139	0.121	0.038

Table 3.1: d_N/d_S is lower in Spike than RdRp. A single d_N/d_S value was computed for gene (or spike domain) and each HCoV using MEME

3.2.4 Rate of Adaptation in RdRp and subdomains of spike

Therefore, as a complement to the divergence analysis, we implemented an alternative to the d_N/d_S method that was specifically designed to detect positive selection within RNA virus populations (Bhatt et al., 2011). Compared with traditional d_N/d_S methods, the Bhatt method has the advantages of: 1) measuring the strength of positive selection within a population given sequences collected over time, 2) higher sensitivity to identifying mutations that occur only once and sweep through the population, and 3) correcting for deleterious mutations (Bhatt et al., 2010, 2011). Briefly, this method defines a class of neutrally-evolving nucleotide sites as those with synonymous mutations or where nonsynonymous polymorphisms occur at medium frequency. Then, the number of fixed and high-frequency nonsynonymous sites that exceed the neutral expectation are calculated. This method compares nucleotide sequences at each timepoint (the ingroup) to the consensus nucleotide sequence at the first time point (the outgroup) and yields an estimate of the number of adaptive substitutions within a given genomic region at each of these timepoints.

We adapted this method to detect adaptive substitutions in seasonal HCoVs. As shown in Figure 3.4, OC43 lineage A has continuously amassed adaptive substitutions in spike over the past >30 years while RdRp has accrued few, if any, adaptive substitutions. These adaptive substitutions are located within the S1, and not the S2, domain of spike (Figure

3.4). We observe a largely linear accumulation of adaptive substitutions in spike and S1 through time, although the method does not dictate a linear increase. This observation suggests that spike (and S1 in particular) is evolving in response to a continuous selective pressure. This is exactly what would be expected if these adaptive substitutions are evidence of antigenic evolution resulting from an evolutionary arms race between spike and the host immune system.

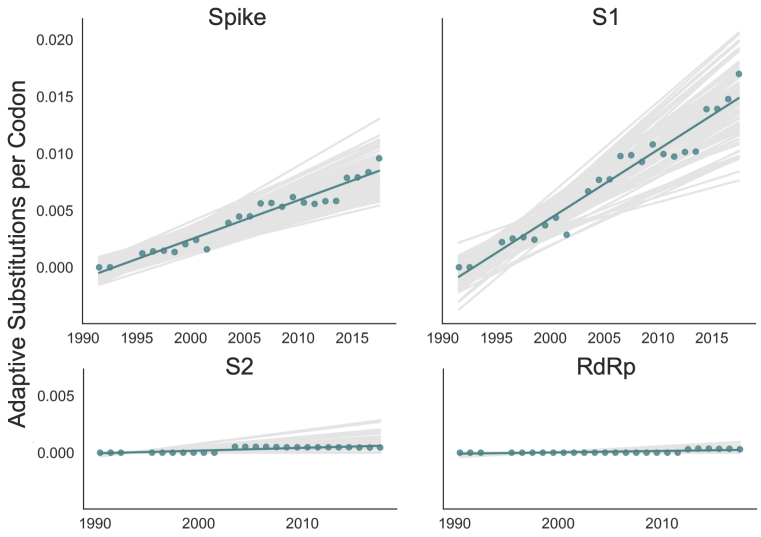

Figure 3.4: Adaptive substitutions accumulate over time in OC43 lineage A spike S1. Adaptive substitutions per codon within OC43 lineage A spike, S1, S2 and RdRp as calculated by our implementation of the Bhatt method. Adaptive substitutions are computed in sliding 3-year windows, and only for timepoints that contain 3 or more sequenced isolates. Red dots display estimated values calculated from the empirical data and red lines show linear regression fit to these points. Grey lines show the distribution of regressions fit to the computed number of adaptive substitutions from 100 bootstrapped datasets. Year is shown on the x-axis.

We estimate that OC43 lineage A accumulates roughly 0.61×10^{-3} adaptive substitutions per codon per year (or 0.45 adaptive amino acid substitutions in S1 each year) in the S1 domain of spike, while the rate of adaptation in OC43 lineage B is slightly higher and is estimated to result in an average 0.56 adaptive substitutions in S1 per year (Figure 3.5). The S1 domain of 229E is estimated to accrue 0.26 adaptive substitutions per year (a rate of 0.47×10^{-3} adaptive substitutions per codon per year).

A benefit of the Bhatt method is the ability to calculate the strength of selection, which allows us to compare these seasonal HCoVs to other viruses. We used our implementation

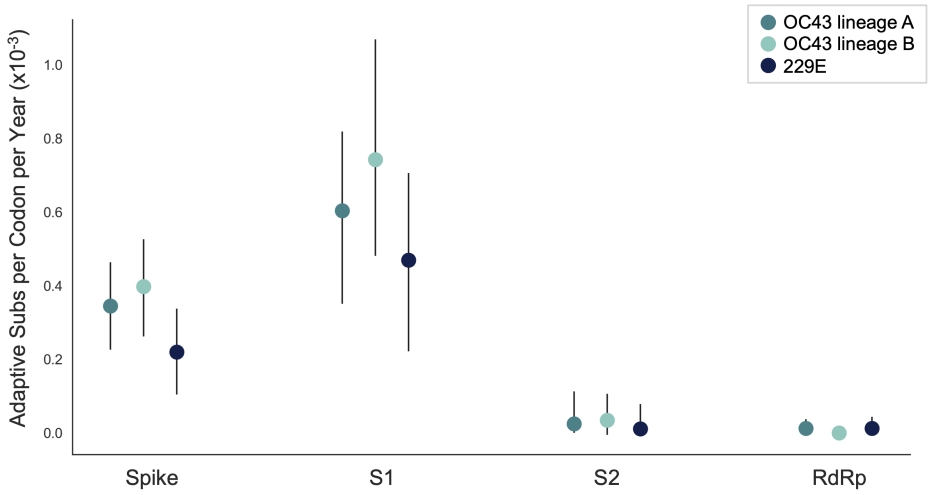

Figure 3.5: The rate of adaptive substitution is highest in spike S1. Adaptive substitutions per codon per year as calculated by our implementation of the Bhatt method. Rates are calculated within Spike, S1, S2 and RdRp for 229E and OC43 lineages. Error bars show 95% bootstrap percentiles from 100 bootstrapped datasets.

of the Bhatt method to calculate the rate of adaptation for influenza A/H3N2, which is known to undergo rapid antigenic evolution (Rambaut et al., 2008; Yang and Nielsen, 2000), measles, which does not (Fulton et al., 2015), and influenza B strains Vic and Yam, which evolve antigenically at a slower rate than A/H3N2 (Bedford et al., 2014). We estimate that the receptor-binding domain of influenza A/H3N2 accumulates adaptive substitutions between 2 and 3 times faster than the HCoVs OC43 and 229E (Figure 3.6). The rate of adaptive substitution in influenza B/Yam and B/Vic are on par with the seasonal HCoVs. We detect no adaptive substitutions in the measles receptor-binding protein. These results put the evolution of the S1 domain of OC43 and 229E in context, indicating that the S1 domain is under positive selection, and that this positive selection generates new variants in the putative antigenic regions of these HCoVs at about the same rate as influenza B strains and about half the rate of the canonical example of antigenic evolution, the HA1 domain of influenza A/H3N2.

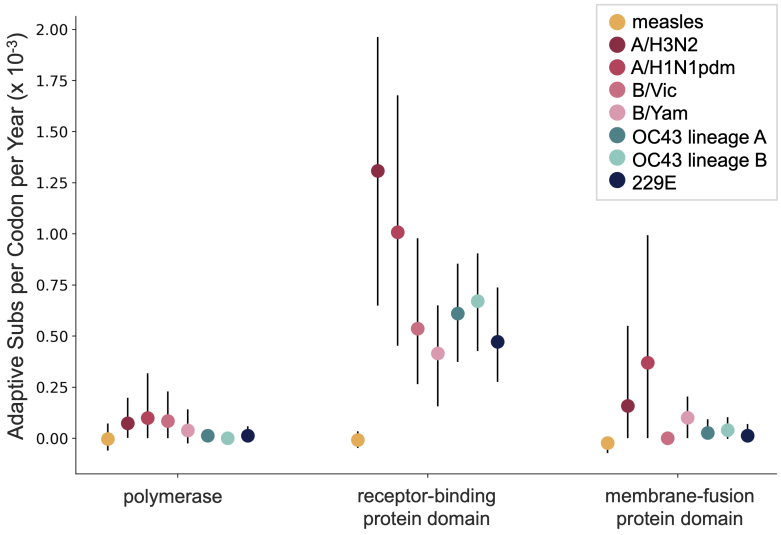

Figure 3.6: OC43 and 229E spike S1 accumulates adaptive substitutions faster than measles but slower than influenza A/H3N2. Comparison of adaptive substitutions per codon per year between measles (yellow), 4 influenza strains (A/H3N2, A/H1N1pdm, B/Vic and B/Yam- shown in shades of red), OC43 lineage A (dark teal), OC43 lineage B (light teal), and 229E (dark blue). The polymerase, receptor binding domain and membrane fusion domain for influenza strains are PB1, HA1 and HA2. For both HCoVs, they are RdRp, S1 and S2, respectively. For measles, the polymerase is the L gene, the receptor-binding protein is the H gene and the fusion protein is the F gene. Error bars show 95% bootstrap percentiles from 100 bootstrapped datasets.

3.2.5 Validation that rate of adaptation is not biased by recombination

Because coronaviruses are known to recombine, and recombination has the potential to impact evolutionary analyses of selection, we sought to verify that our results are not swayed by the presence of recombination. To do this, we simulated the evolution of OC43 lineage A spike and RdRp genes under varying levels of recombination and positive selection (representative phylogenies of simulated spike evolution can be seen in Figure 3.S8) and used our implementation of the Bhatt method to identify adaptive evolution. As the strength of positive selection increases, we detect a higher rate of adaptive evolution, regardless of the level of recombination (Figure 3.7). This demonstrates that our estimates of adaptive evolution are not biased by recombination events.

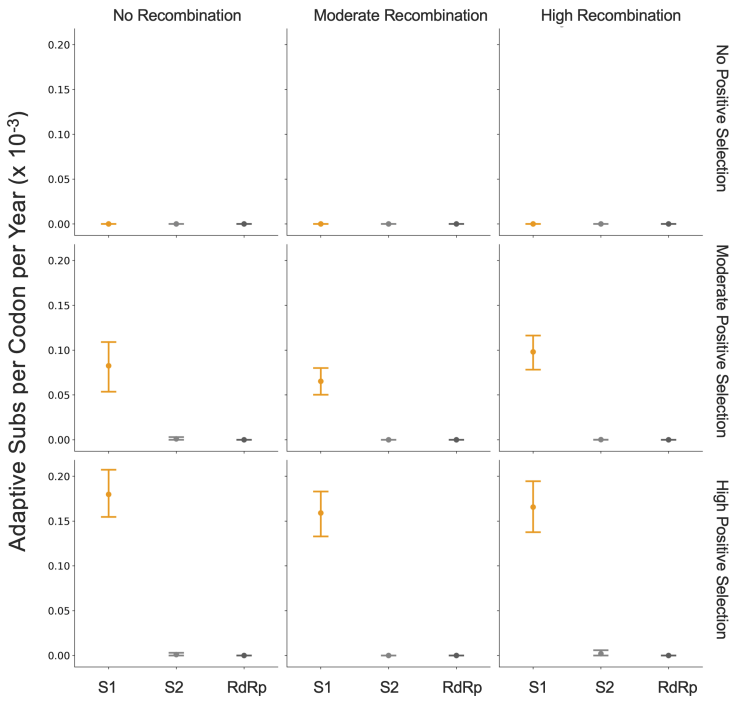

Figure 3.7: Detection of positive selection is not biased by recombination. OC43 lineage A sequences were simulated with varying levels of recombination and positive selection. The Bhatt method was used to calculate the rate of adaptive substitutions per codon per year for S1 (light orange), S2 (light gray) and RdRp (dark gray). The mean and 95% confidence interval of 10 independent simulations is plotted.

3.2.6 Time to Most Recent Common Ancestor (TMRCA) of RdRp and subdomains of spike

Finally, we know that strong directional selection skews the shape of phylogenies (Volz et al., 2013). In influenza H3N2, immune selection causes the genealogy to adopt a ladder-like shape where the rungs are formed by viral diversification and each step is created by the appearance of new, antigenically-superior variants that replace previous variants. This ladder-like shape can also be seen in the phylogenies of the OC43 and 229E (Figure 3.1). In this case, selection can be quantified by the timescale of population turnover as measured by the Time to Most Recent Common Ancestor (TMRCA), with the expectation that stronger selection will result in more frequent steps and therefore a smaller TMRCA measure (Bedford et al., 2011). We computed average TMRCA values from phylogenies built on Spike, S1, S2 or RdRp sequences of OC43 lineage A and 229E (Table 3.2). We did not compute TMRCA for OC43 lineage B because the limited number of available RdRp sequences for this lineage

mean that TMRCA can only be calculated for about 4 years, which could artificially skew the TMRCA estimates. Our estimates of HCoV spike TMRCA are roughly 2-2.5 longer the estimated TMRCA for influenza A/H3N2 hemagglutinin (Bedford et al., 2011).

We observe that, for both OC43 lineage A and 229E, the average TMRCA is lower in spike than RdRp and lower in S1 versus S2. These results suggest strong directional selection in S1, likely driven by pressures to evade the humoral immune system. The difference in TMRCA between S1 and S2 is indicative not only of differing selective pressures acting on these two spike domains, but also of intra-spike recombination. This is because the immune selection imposed on S1, should also propagate neutral hitch-hiker mutations in closely-linked regions such as S2. The difference in TMRCA suggests that recombination may uncouple these regions. Recombination can also push TMRCA to higher values, though this should not have a larger impact on RdRp than S1. The contributions of the forces of directional selection and recombination are difficult to parse from the TMRCA results. This emphasizes the importance of using methods, such as the Bhatt method, that are robust to recombination to detect adaptive evolution.

	Spike	S1	S2	RdRp
OC43 lineage A	4.67 (4.04, 5.28)	3.45 (2.86, 4.05)	13.05 (11.24, 14.97)	17.39 (15.63, 19.15)
229E	4.19 (3.13, 5.25)	2.23 (1.76, 2.69)	5.08 (3.93, 6.23)	4.86 (4.04, 5.69)

Table 3.2: Mean TMRCA is lower in S1 than RdRp or S2. Average TMRCA values (in years) for OC43 lineage A and 229E. The 95% confidence intervals are indicated in parentheses below mean TMRCA values.

3.2.7 Application of methods for identifying adaptive evolution to HKU1 and NL63

Because HKU1 was identified in the early 2000's, there are fewer longitudinally-sequenced isolates available for this HCoV compared to 229E and OC43 (Figure 3.S2). Consequently, the phylogenetic reconstructions and divergence analysis of HKU1 have a higher level of

uncertainty. To begin with, it is less clear from the phylogenies whether HKU1 represents a single HCoV lineage like 229E or, instead, should be split into multiple lineages like OC43 (Figure 3.1). Because of this, we completed all antigenic analyses for HKU1 twice: once considering all isolates to be members of a single lineage, and again after splitting isolates into 2 separate lineages. These lineages are arbitrarily labeled 'A' and 'B' as was done for OC43. When HKU1 is considered to consist of just one lineage, there is no signal of antigenic evolution by divergence analysis (Figure 3.S4B) or by the Bhatt method of estimating adaptive evolution (Figure 3.S6A). However, when HKU1 is assumed to consist of 2 co-circulating lineages, HKU1 lineage A has a markedly higher rate of adaptive substitutions in S1 than in S2 or RdRp (Figure 3.S6B).

To demonstrate the importance of having a well-sampled longitudinal series of sequenced isolates for our antigenic analyses, we returned to our simulated OC43 S1 datasets. We mimicked shorter longitudinal series by truncating the dataset to only 24, 14, 10, or 7 years of samples and ran the Bhatt analysis on these sequentially shorter time series (Figure 3.S7). The rates of adaptation estimated from the truncated datasets can be compared to the "true" rate of adaptation calculated from all simulated data. This simulated data reveals a general trend that less longitudinal data reduces the ability to detect adaptive evolution by skewing the estimated rate away from the "truth" and increasing the uncertainty of the analysis. Given the dearth of longitudinal data for HKU1, we do not feel that it is appropriate to make strong conclusions about adaptive evolution, or lack thereof, in this HCoV.

Despite being identified at roughly the same time as HKU1, substantially more NL63 isolates have been sequenced (Figure 3.S2) making the phylogenetic reconstruction and evolutionary analyses of this virus correspondingly more reliable. We do not observe evidence for adaptive evolution in NL63 (Figure 3.S4A and Figure 3.S6A) and this lack of support for adaptive evolution in the NL63 spike gene is more likely to reflect an actual lack of adaptive evolution in this virus.

3.3 Discussion

Using several corroborating methods, we provide evidence that the seasonal HCoVs OC43 and 229E undergo adaptive evolution in S1, the region of the spike protein exposed to human humoral immunity (Figures 3, 4 and 5). We additionally confirm that RdRp and S2 do not show signals of adaptive evolution. We observe that S1 accumulates between 0.3 (229E) and 0.5 (OC43) adaptive substitutions per year. We infer that these viruses accumulate adaptive substitutions at roughly half the rate of influenza A/H3N2 and at a similar rate to influenza B viruses (Figure 3.6). The most parsimonious explanation for the observation of substantial adaptive evolution in S1 is that antigenic drift is occurring in which mutations that escape from human population immunity are selectively favored in the viral population leading to repeated adaptive changes. However, it is formally possible that the adaptive evolution we detect is a result of selective pressures other than evasion of the adaptive immune system. Showing that this is truly antigenic evolution could involve a serological comparison of isolates that differ at S1 residues under positive selection.

In seasonal influenza and measles, the rates of adaptive evolution we estimate correlate well with relative rates of antigenic drift reported by other groups (Fulton et al., 2015; Bedford et al., 2014). The relative rates of adaptation we calculate also match the relative frequency of vaccine strain updates, as would be expected since vaccines must be updated to match antigenically-evolving viruses. Since 2006, the A/H3N2 component of the seasonal influenza vaccine has been updated 10 times (11 different A/H3N2 strains), 4 different B/Vic strains and 4 different B/Yam strains have been included in the vaccine, and the measles vaccine strain has not changed (Global Influenza Surveillance and Response System (GISRS), `https://www.who.int/influenza/vaccines/virus/en/`). Using these numbers as guidance, our results suggest that a vaccine against OC43 or 229E might need to be updated as frequently as the B/Vic and B/Yam components of the influenza vaccine are.

We do not observe evidence of antigenic evolution in NL63 (Figure 3.S4 and Figure 3.S6). This likely represents a lack of marked adaptive evolution in S1. Our finding fits

with a study of NL63 in Kenya, which identified multiple genotypes of NL63 and show that people regularly become reinfected with the same genotype of NL63 (Kiyuka et al., 2018). Additionally, Kiyuka et al found that these genotypes circulate locally for a long period of time, suggesting a decent amount of viral diversity and a potential lack of evolution due to immune selection. Though our results cannot explain why OC43 and 229E likely evolve antigenically while NL63 does not, Kiyuka et al observe that NL63 reinfections are sometimes enhanced by a previous infection and hypothesize that NL63 is actually under purifying selection at epitope sites (Kiyuka et al., 2018).

Though analysis of all HCoVs would benefit from more sequenced isolates, there is substantially less longitudinal sequencing data available for HKU1. Thus, despite finding no evidence of antigenic evolution in HKU1 (Figure 3.S4 and Figure 3.S6), it is possible that a more completely sampled time series of HKU1 genome sequences could alter the result for this virus (Figure 3.S7).

Our conclusions of adaptive evolution in S1, arrived at through computational analyses of sequencing data, agree with studies that observe reinfection of subjects by heterologous isolates of 229E (Reed, 1984), sequential dominance of specific genotypes of OC43 (Lau et al., 2011; Zhang et al., 2015), and common reinfection by seasonal HCoVs from longitudinal serological data (Edridge et al., 2020). In this latter study, HCoV infections were identified from longitudinal serum samples by assaying for increases in antibodies against the nucleocapsid (N) protein of representative OC43, 229E, HKU1, and NL63 viruses. This study concluded that the average time between infections was 1.5–2.5 years, depending on the HCoV (Edridge et al., 2020). In comparison, influenza H3N2 reinfects people roughly every 5 years (Kucharski et al., 2015). Thus, frequent reinfection by seasonal HCoVs is likely due to a combination of factors and suggests waning immune memory, and/or incomplete immunity against reinfection, in addition to antigenic drift.

Human coronaviruses are a diverse grouping split, phylogenetically, into two genera: NL63 and 229E are alphacoronaviruses, while OC43, HKU1, MERS, SARS, and SARS-CoV-

2 are betacoronaviruses. The method of cell-entry does not seem to correlate with genus. Coronaviruses bind to a remarkable range of host-cell receptors including peptidases, cell adhesion molecules and sugars. Amongst the seasonal HCoVs, OC43 and HKU1 both bind 9-O-acetylsialic acid (Hulswit et al., 2019), while 229E binds human aminopeptidase N (hAPN) and NL63 binds angiotensin-converting enzyme 2 (ACE2) (Liu et al., 2020a). Despite a relatively large phylogenetic distance and divergent S1 structures, NL63 and SARS-CoV-1 and SARS-CoV-2 bind to the same host receptor using the same virus-binding motifs (VBMs) (Li, 2016). This VBM is located in the C-terminal domain of S1 (S1-CTD), which fits within the trend of S1-CTD receptor-binding in CoVs that bind protein receptors (Hofmann et al., 2006; Li, 2016). This is opposed to the trend amongst CoVs that bind sugar receptors, where receptor-binding is located within the S1 N-terminal domain (S1-NTD) (Li, 2016). This localization roughly aligns with our observations that the majority of the repeatedly-mutated sites occur toward the C-terminal end of 229E S1 and the N-terminal end of OC43 S1 (Figure 3.2).

Here, we have provided support that at least 2 of the 4 seasonal HCoVs evolve adaptively in the region of spike that is known to interact with the humoral immune system. These two viruses span both genera of HCoVs, though due to the complexity of HCoV receptor-binding and pathology mentioned above, it is not clear whether or not this suggests that other HCoVs, such as SARS-CoV-2, will also evolve adaptively in S1. This is important because, at the time of writing of this manuscript, many SARS-CoV-2 vaccines are in production and most of these exclusively include spike (Krammer, 2020). If SARS-CoV-2 evolves adaptively in S1 as the closely-related HCoV OC43 does, it is possible that the SARS-CoV-2 vaccine would need to be frequently reformulated to match the circulating strains, as is done for seasonal influenza vaccines.

3.4 Methods

All data, source code and analyses can be found at `https://github.com/blab/seasonal-cov-adaptive-evolution`. All phylogenetic trees constructed and analyzed in this manuscript can be viewed interactively at `https://nextstrain.org/community/blab/seasonal-cov-adaptive-evolution`. All analysis code is written in Python 3 (Python Programming Language, SCR_008394) in Jupyter notebooks (Jupyter-console, RRID:SRC_018414).

3.4.1 Sequence data

All viral sequences are publicly accessible and were downloaded from ViPR (`www.viprbrc.org`) under the "Coronavirdiae" with host "human" (Pickett et al. 2012). Sequences labeled as "OC43", "229E", "HKU1" and "NL63" were pulled out of the downloaded FASTA file into 4 separate data files. Additionally, a phylogeny of all downloaded human coronaviruses was made and unlabeled isolates that clustered within clades formed by labeled OC43, 229E, HKU1 or NL63 isolates were marked as belonging to that HCoV type and added to our data files. Code for these data-parsing steps is located in `data-wrangling/postdownload_formatting_for_rerun.ipynb`.

3.4.2 Phylogenetic inference

For each of the 4 HCoV datasets, full-length sequences were aligned to a reference genome using the augur align command (Hadfield et al., 2018) and MAFFT (Katoh et al., 2002). Individual gene sequences were then extracted from these alignments if sequencing covered 50% or more of the gene using the code in `data-wrangling/postdownload_formatting_for_rerun.ipynb`. Sequence files for each gene are located in the data/ directory within each HCoV parent directory (ex: `oc43/data/oc43_spike.fasta`). A Snakemake file (Köster and Rahmann, 2012) within each HCoV directory follows the general outline of a Nextstrain build (Nextstrain,

RRID:SCR_018223) and was used to align each gene to a reference strain and build a time-resolved phylogeny with IQ-Tree v1 (Nguyen et al., 2015) and TimeTree (Sagulenko et al., 2018). Phylogenies were viewed to identify the distribution of genotypes throughout the tree, different lineages, and signals of recombination using the nextstrain view command (Hadfield et al., 2018). The clock rate of the phylogeny based on spike sequences for each isolate (as shown in Figure 3.1 and Figure 3.S2) was 0.0005 substitutions per nucleotide site per year for OC43, 0.0006 for 229E, 0.0007 for NL63, and 0.0062 for HKU1. All NL63 and HKU1 trees were rooted on an outgroup sequence. For NL63, the outgroup was 229e/AF304460/229e_ref/Germany/2000 and for HKU1 the outgroup was mhv/NC_048217_1/mhv/2006. Clock rates for the phylogenies built on each individual gene can be found within the `results/` directory within each HCoV parent directory (ex: `oc43/results/branch_lengths_oc43_spike.json`).

3.4.3 Mutation counting

Amino acid substitutions at each position in spike were tallied from the phylogeny. In other words, the phylogenetic reconstruction of spike sequences returns nucleotides changes to the ancestral sequence along each branch. The number of times this changed amino acid identity at each position was tallied. This analysis was conducted using code in `antigenic_evolution/site_mutation_rank.ipynb`.

3.4.4 Divergence analysis

For each HCoV lineage and each gene, synonymous and nonsynonymous divergence was calculated at all timepoints as the average Hamming distance between each sequenced isolate and the consensus sequence at the first timepoint (founder sequence). The total number of observed differences between the isolate and founder nucleotide sequences that result in nonsynonymous (or synonymous) substitutions is divided by the number of possible nucleotide mutations that result in nonsynonymous (or synonymous) substitutions, weighted by kappa,

to yield an estimate of divergence. Kappa is the ratio of rates of transitions:transversions, and was calculated by averaging values from spike and RdRp trees built by BEAST 2.6.3 (Bouckaert et al., 2019) using the HKY+gamma4 model with 2 partitions and "coalescent constant population". All BEAST results are found in .log files in gene- and HCoV-specific subdirectories within `beast/` . Divergence is calculated from nucleotide alignments. Sliding 3-year windows were used and only timepoints that contained at least 2 sequences were considered. The concept for this analysis is from (Zanini et al., 2015) and code for our adaptation is `in antigenic_evolution/divergence_weighted.ipynb`. The ratios of divergence shown in Figure 3.S5 are also calculated in this notebook.

3.4.5 Calculation of $\mathrm{d_N/d_S}$

A d_N/d_S value was calculated for RdRp, spike, S1 and S2 of each HCoV using the Datamonkey (Weaver et al., 2018) implementation of MEME (Mixed Effects Model of Evolution) (Murrell et al., 2012). Aligned FASTA files (ex: `oc43/results/aligned_oc43_rdrp.fasta`) were uploaded to Datamonkey (`http://datamonkey.org/meme`) and d_N/d_S value was recorded as the calculated Global MG94xREV model nonsynonymous/synonymous rate ratio.

3.4.6 Implementation of the Bhatt method

The rate of adaptive evolution was computed using an adaptation of the Bhatt method (Bhatt et al., 2010, 2011). For each lineage and each genomic region, we partitioned all available sequences into sliding 3-year windows and only used timepoints that contained at least 3 sequences in the analysis. We compared nucleotide sequences at each timepoint (the ingroup) to the consensus nucleotide sequence at the first time point (the outgroup). Eight estimators (silent fixed, replacement fixed, silent high frequency, replacement high frequency, silent mid-frequency, replacement mid-frequency, silent low frequency and replacement low-frequency) are calculated by the site-counting method (Bhatt et al., 2010). In the site-

counting method, each estimator is the product of the fixation or polymorphism score times the silent or replacement score, summed for each site in that frequency class. Fixation and polymorphism scores depend on the number of different nucleotides observed at the site and whether the outgroup base is present in the ingroup. Selectively neutral sites are assumed to contain the classes of silent polymorphisms and replacement polymorphisms occurring at a frequency between 0.15 and 0.75. A class of nonneutral, adaptive sites is then identified as having an excess of replacement fixations or polymorphisms (Bhatt et al., 2011). For each lineage and gene, 100 bootstrap alignments and ancestral sequences were generated and run through the Bhatt method to assess the statistical uncertainty of our estimates of rates of adaptation (Bhatt et al., 2011). The rate of adaptation (per codon per year) shown in Figure 3.5 is calculated by linear regression of the time series values of adaptive substitutions per codon (Figure 3.4). Our code for implementing the Bhatt method is at `antigenic_evolution/bhatt_bootstrapping.ipynb`.

3.4.7 Estimation of rates of adaptation of measles and influenza viruses

Influenza and measles alignments were generated by running Nextstrain the respective Nextstrain builds from `https://github.com/nextstrain/seasonal-flu` and `https://github.com/nextstrain/measles` (Hadfield et al., 2018). The seasonal influenza build was run with 20 year resolution for H3N2, H1N1pdm, Vic and Yam. The rates of adaptation of different genes was calculated using our implementation of the Bhatt method described above. The receptor-binding domain used for influenza was HA1, for measles was the H protein, and for the HCoVs was S1. The membrane fusion protein used for influenza was HA2, for measles was the F protein, and for the HCoVs was S2. The polymerase for influenza was PB1, for measles was the L protein, and for the HCoVs was RdRp (Nsp12). Our code for this analysis is at `antigenic_evolution/bhatt_nextstrain.ipynb`.

3.4.8 Simulation of evolving OC43 sequences

The evolution of OC43 lineage A Spike and RdRp genes was simulated using SANTA-SIM (Jariani et al., 2019). The OC43 lineage A root sequence was used as a starting point and the simulation was run for 500 generations and 10 simulated sequences were sampled every 50 generations. The spike and RdRp genes were simulated separately. Purifying selection was simulated across both genes. Evolution was simulated in the absence of recombination and with moderate and high levels of recombination during replication. Under each of these recombination paradigms, we simulated evolution in the absence of positive selection within spike and with moderate and high levels of positive selection. Positive selection was simulated through exposure-dependent selection at a subset of spike S1 sites proportional to the number of epitope sites in H3N2 HA (Luksza and Lässig, 2014). The simulated selection allows mutations in these "epitope" sites to rise in frequency while also encouraging "epitopes" to change over time (to mimic antigenic novelty). All simulations were run with a nucleotide mutation rate of 1x10-4 (Vijgen et al., 2005). Config files, results and source code for these simulations can be at `santa-sim_oc43a/` and the Bhatt method is implemented on the simulated data in `antigenic_evolution/bhatt_simulated_oc43_data.ipynb`.

3.4.9 Estimation of TMRCA

Mean TMRCA values were estimated for each gene and each HCoV using PACT (Bedford et al., 2011). Briefly, PACT computes TMRCA values by creating a series of subtrees that include only tips positioned within a temporal slice of the full tree and finding the common ancestor of these tips. The overall mean and 95% confidence interval were calculated from the list of TMRCA values in these time slices. The PACT config files and results for each run are in the directory `antigenic_evolution/pact/`. The TMRCA estimations and subsequent analyses are executed by code in `antigenic_evolution/tmrca_pact.ipynb`.

3.5 Supplemental Information

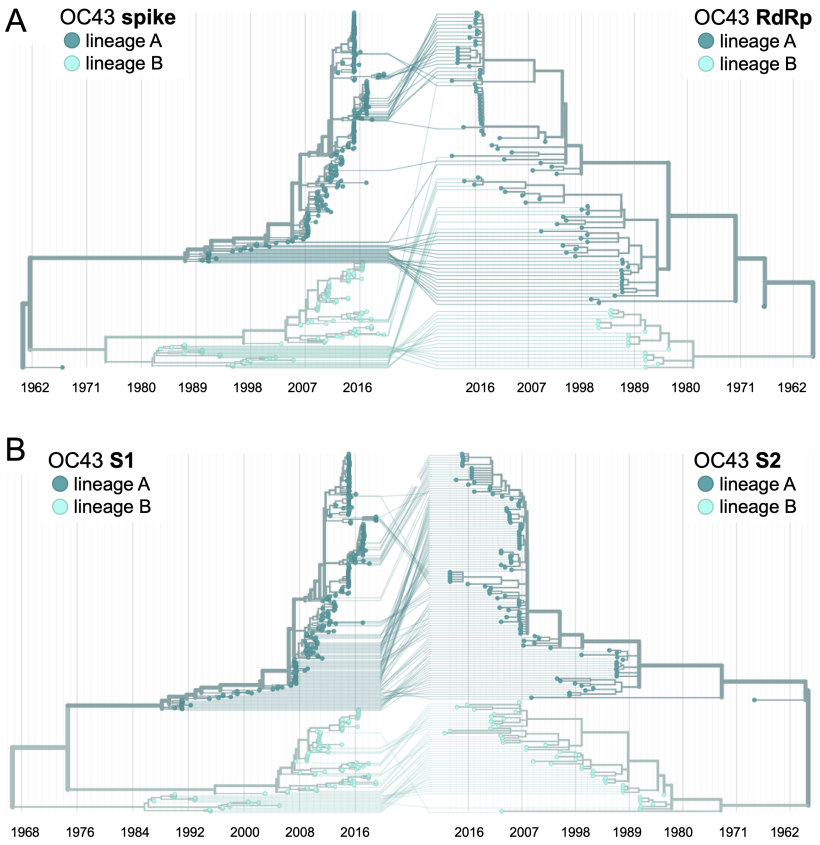

Figure 3.S1: Recombination occurs between HCoV isolates. A tanglegram draws lines between an isolate's position on two phylogenies built on different genes (or genomic regions). Dramatic differences in an isolate's position on one tree versus another is indicative of recombination. **A)** Phylogenetic relationships between OC43 isolates based on spike sequences (left) versus relationships based on RdRp sequences (right). Light teal lines that connect isolates classified as lineage A based on their RdRp sequence to isolates classified as lineage B based on their spike sequence suggest that recombination occurred in these isolates or their ancestors. **B)** Phylogenetic reconstruction of OC43 isolates based on S1 sequences (left) versus S2 sequences (right). Year is shown on the x-axis.

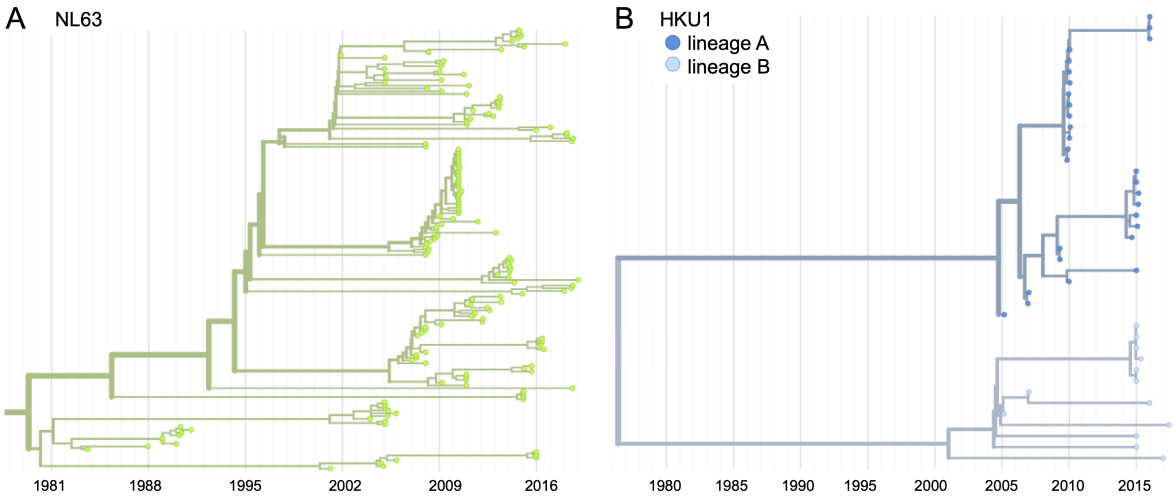

Figure 3.S2: Phylogenetic trees for seasonal HCoVs NL63 and HKU1. Phylogenies built from **A)** NL63 spike sequences from 159 isolates over 37 years, and **B)** HKU1 spike sequences from 41 isolates over 13 years. HKU1 bifurcate immediately after the root and is split into lineage A (darker blue) and lineage B (lighter blue). NL63 contains just one lineage (green). Both HCoVs are rooted on an outgroup sequence. For the analyses in this paper, the evolution of each gene (or genomic region) is considered separately, so phylogenies are built for each viral gene and those phylogenies are used to split isolates into lineages for each gene. These are temporally resolved phylogenies with year shown on the x-axis. The clock rate of each HCoV is listed in the Methods "Phylogenetic inference" section.

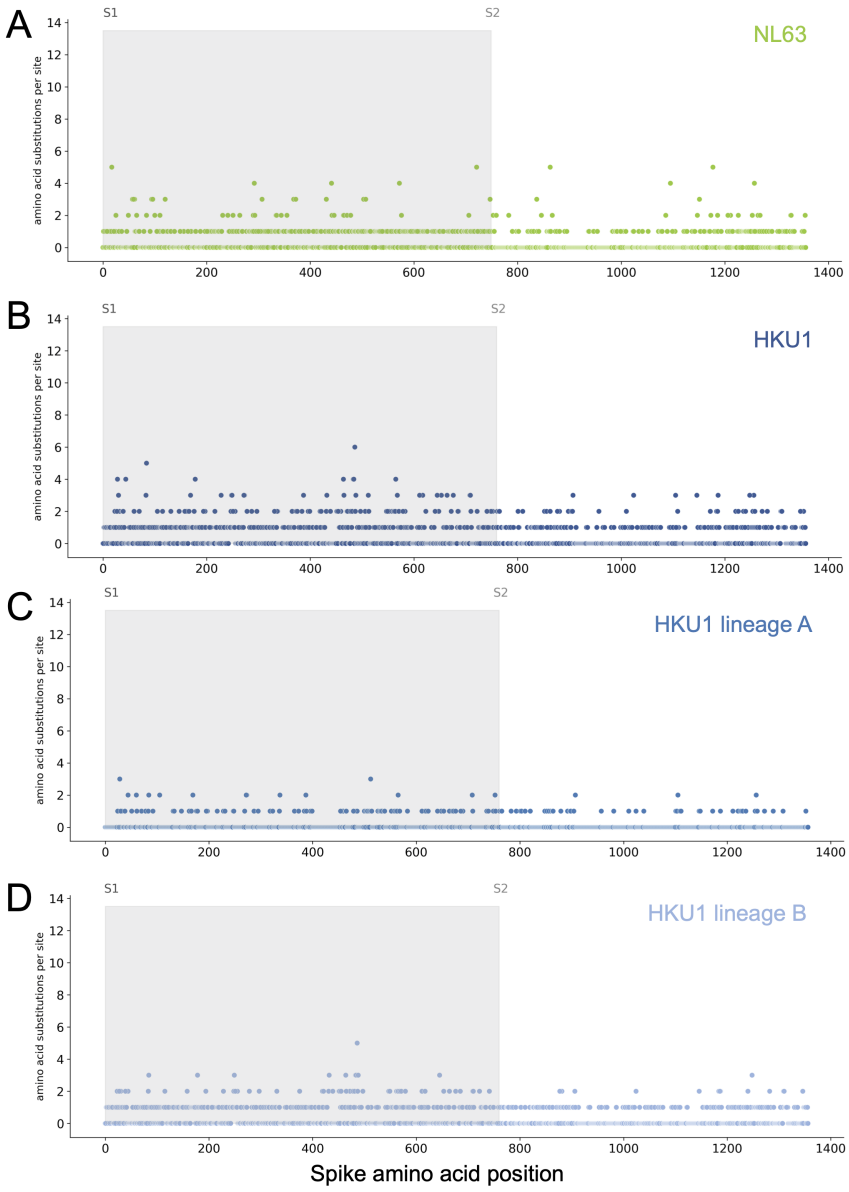

Figure 3.S3: Mutations per at each position within Spike for NL63 and HKU1. Number of substitutions observed at each amino acid position in the spike gene throughout the phylogeny S1 (gray) and S2 (white) are indicated by shading and the number of substitutions per site is indicated by a dot and color-coded by HCoV lineage. **A)** NL63, **B)** HKU1 (assuming all HKU1 isolates are a single lineage), **C)** HKU1 lineage A, **D)** HKU1 lineage B (assuming there are 2 co-circulating HKU1 lineages).

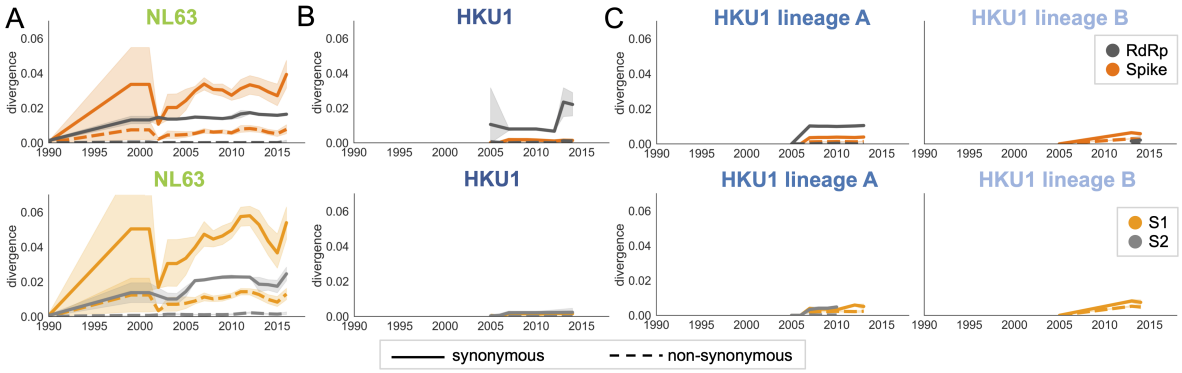

Figure 3.S4: Nonsynonymous divergence in NL63 and HKU1. Nonsynonymous (dashed lines) and synonymous divergence (solid lines) within the spike (dark orange) and RdRp (dark gray) genes and within S1 (light orange) and S2 (light gray) over time. Divergence is the average Hamming distance from the ancestral sequence, computed in sliding 3-year windows which contain at least 2 sequenced isolates. Shaded region shows 95% confidence intervals. **A)** NL63, **B)** HKU1 (assuming all HKU1 isolates belong to a single lineage), and **C)** HKU1 (divided into 2 co-circulating lineages). Year is shown on the x-axis. Note that x- and y-axis scales are shared between the subplots but are different than Figure 3.3.

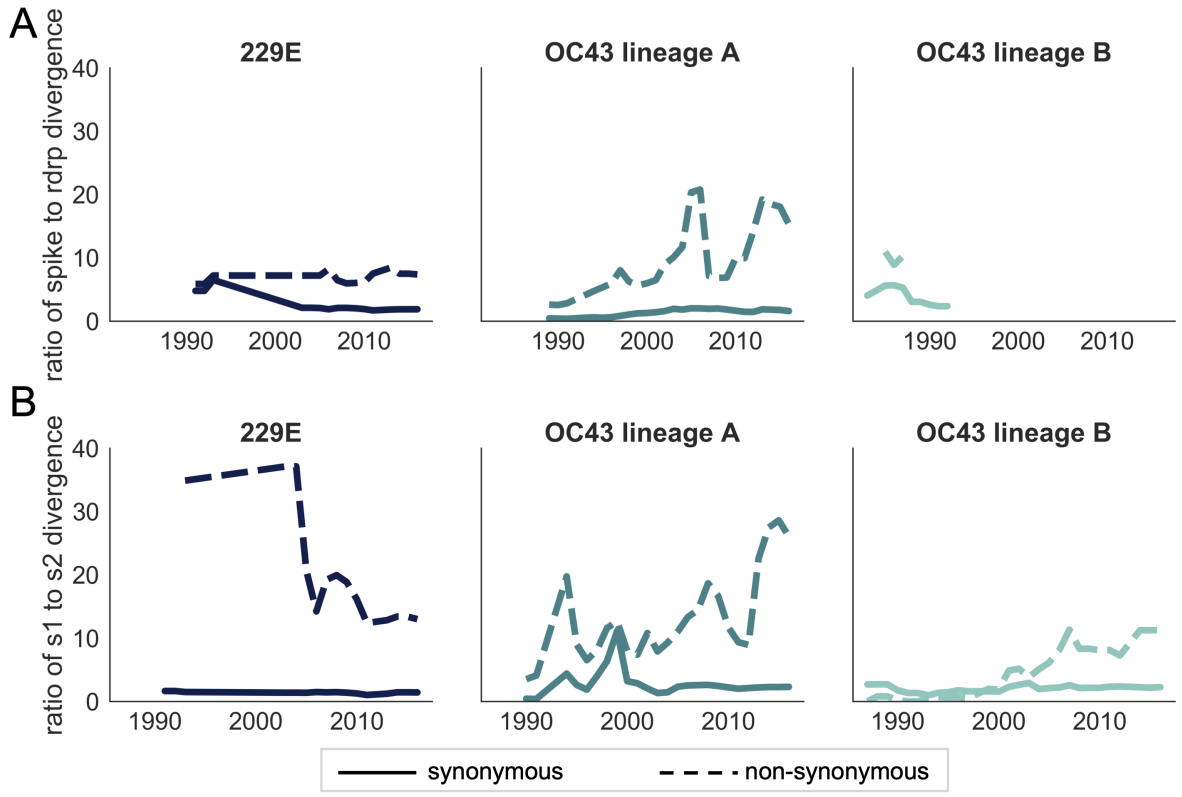

Figure 3.S5: Ratio of divergence between genomic regions. A) the ratio of nonsynonymous divergence in spike to nonsynonymous divergence in RdRp (dashed lines) and the equivalent ratio of synonymous divergence (solid lines) is shown for 229E (dark blue), OC43 lineage A (dark teal), and OC43 lineage B (light teal). **B)** the same ratios of divergence as in panel A, except comparing S1 and S2. Year is on the x-axis.

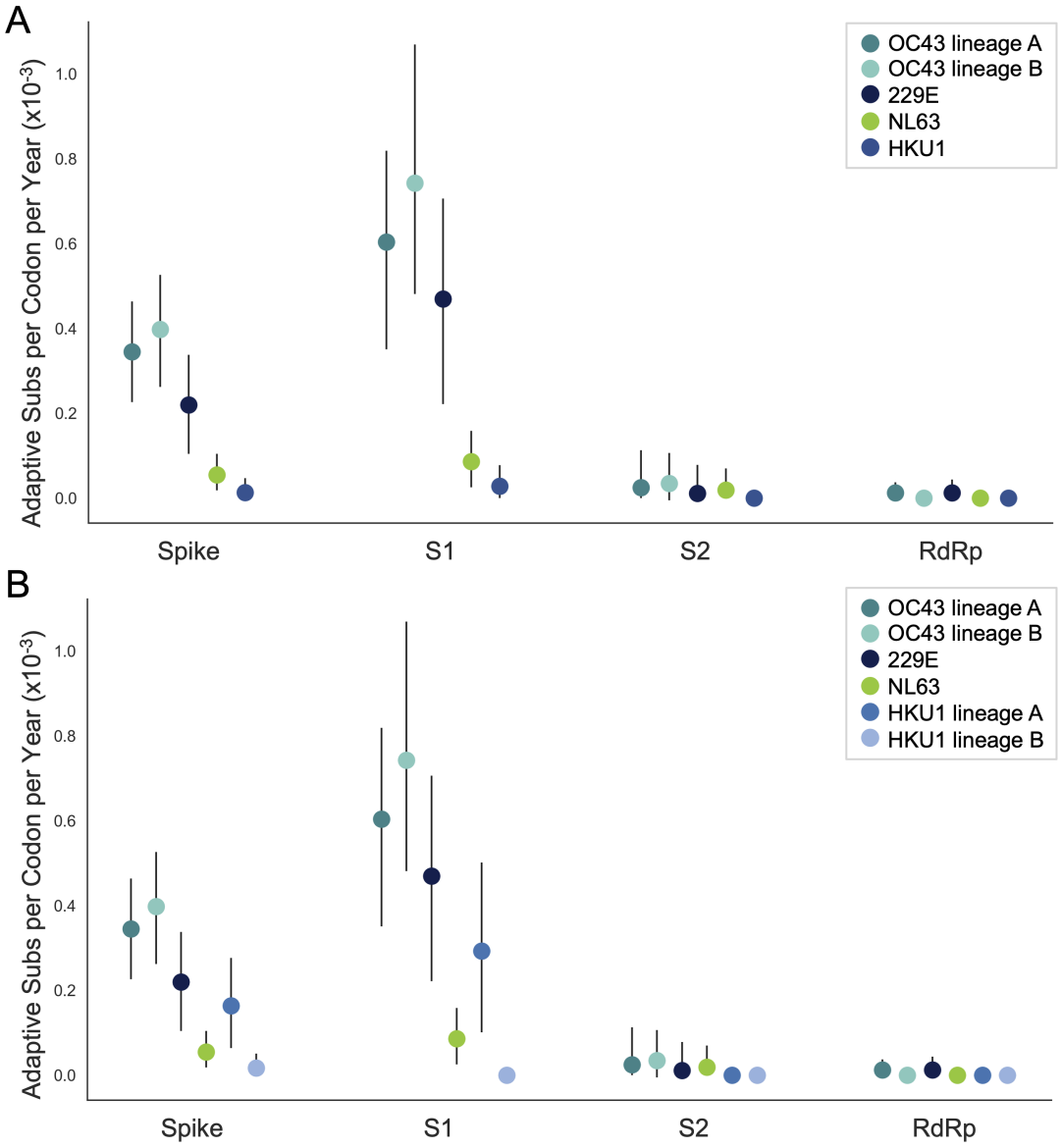

Figure 3.S6: NL63 and HKU1 have low rates of adaptation in spike. As in Figure 3.5, adaptive substitutions per codon per year are calculated by our implementation of the Bhatt method. **A)** NL63 (green) and HKU1 (blue) are both considered to consist of a single lineage. **B)** HKU1 is divided into 2 co-circulating lineages (blue and light blue). The calculated rates of adaptive substitution within spike, S1, S2 and RdRp are plotted alongside 229E and OC43 for comparison. Error bars show 95% bootstrap percentiles from 100 bootstrapped datasets.

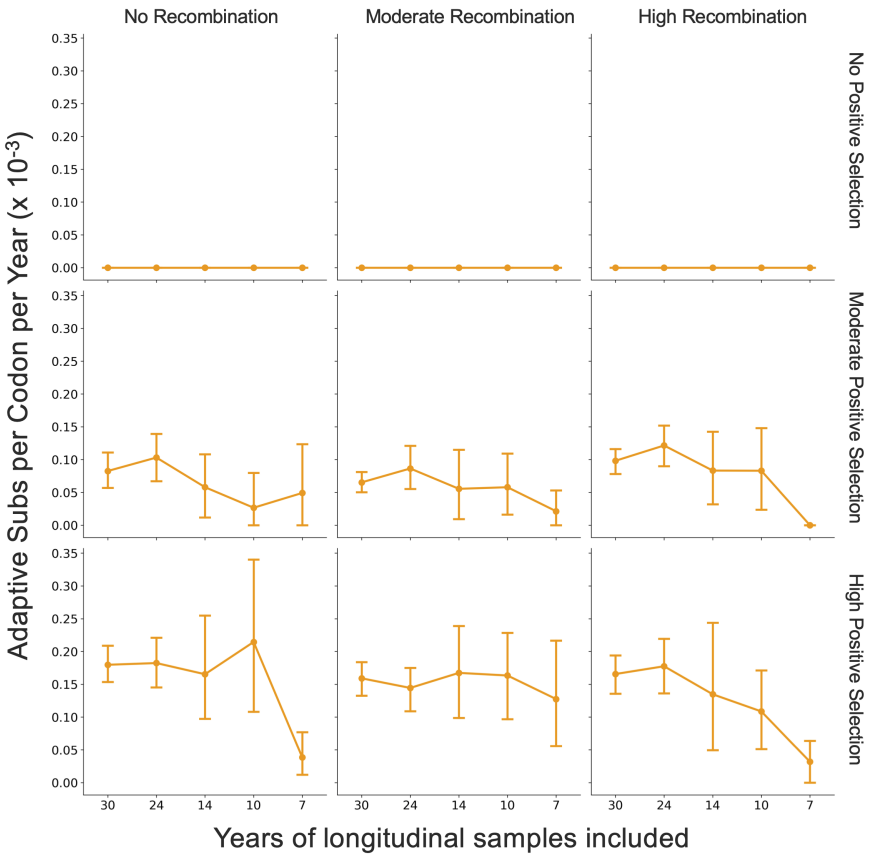

Figure 3.S7: Fewer years of longitudinally-sampled isolates reduces ability to detect rate of adaptation. OC43 lineage A S1 sequences were simulated under conditions of no, moderate and high rates of recombination in combination with no, moderate or high strength of positive selection. The Bhatt method was used to calculate the "true" rate of adaptive evolution under each of these scenarios using all available simulated sequence data (30 years), or the estimated rate if only the most recent 24, 14, 10 or 7 years of simulated sequences were used. The mean and 95% confidence intervals of 10 independent simulations are plotted.

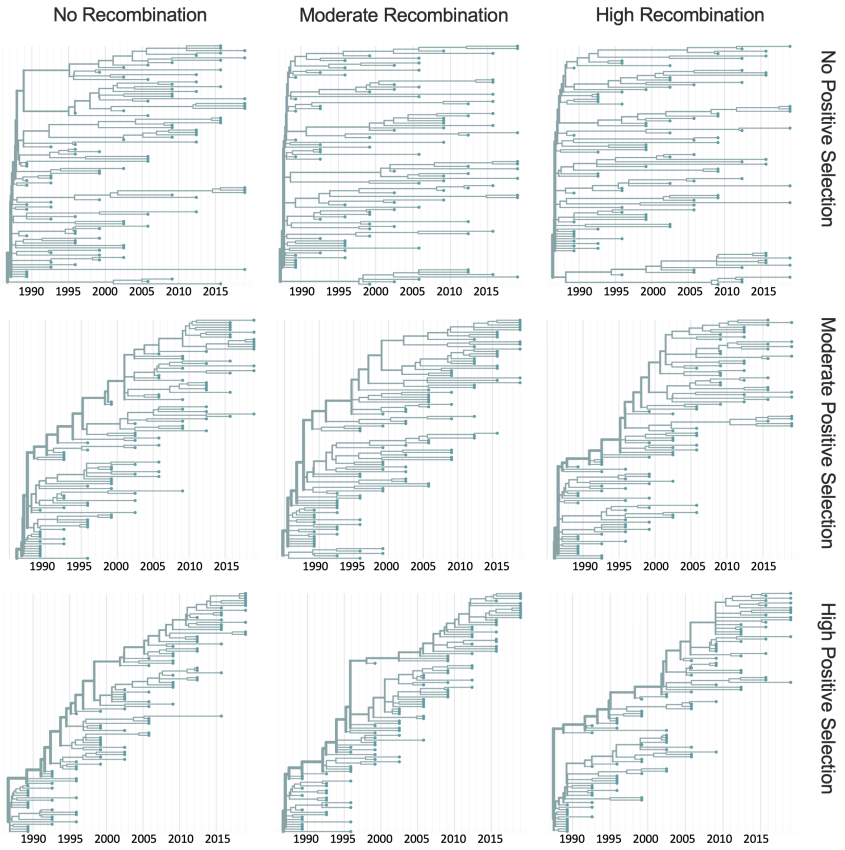

Figure 3.S8: Representative phylogenies of simulated spike data. OC43 lineage A spike sequence evolution was simulated under conditions of no, moderate and high rates of recombination in combination with no, moderate or high strength of positive selection. This figure shows time-resolved phylogenies built from 1 of the 10 independent simulations under each recombination/selection regime.

Chapter 4

Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2

This work was originally published as a preprint with *bioXriv* at `https://doi.org/10.1101/2021.09.11.459844`.

4.1 Introduction

After 20 months of global circulation, basal lineages of SARS-CoV-2 have been almost completely replaced by derived, variant lineages. These lineages are classified by the WHO as variants of concern (VOCs) or variants of interest (VOIs) based on genetic, phenotypic and epidemiological differences (Konings et al., 2021). The effort to track the spread of these variants (and of the pandemic in general) through genomic epidemiology has resulted in a massive corpus of sequenced viral genomes. In the GISAID EpiCoV database alone, there are 2.5 million sequences and counting as of the end of July 2021 (Shu and McCauley, 2017). This thorough sampling offers an opportunity to investigate the evolutionary dynamics of a virus as it entered a naive population, spread rampantly, and, subsequently, began to transmit through previously exposed hosts. Here, we are particularly interested in whether SARS-CoV-2 viruses show phylogenetic evidence of adaptive evolution during the first year and a half of transmission and in the presence of mounting immunity in humans.

Seasonal influenza and seasonal coronaviruses both exhibit continual adaptive evolution during endemic circulation in the human population. In the case of influenza H3N2, transmission through an exposed host population results in adaptive evolution within hemagglutinin (HA). The HA1 subunit of hemagglutinin both mediates binding to host cell receptors and

is the primary target for neutralizing antibodies. Thus, in the context of an exposed host, selection for receptor binding avidity (Hensley et al., 2009) and for escape from humoral immunity (Bedford et al., 2014) drive fixation of mutations in the HA1 subunit. The coronavirus protein subunit equivalent in function to HA1 is spike S1. Previously, we showed that at least two coronaviruses (229E and OC43) exhibit adaptive evolution concentrated in the S1 subunit of spike (Kistler and Bedford, 2021). By demonstrating that strong immune responses to a particular historical isolate of 229E do not neutralize 229E viruses that circulate years afterwards, Eguia et al confirmed that 229E evolves antigenically (Eguia et al., 2021).

Primary methods used to detect adaptive evolution in seasonal influenza and seasonal coronaviruses rely on the fixation (or near fixation) of nonsynonymous changes, and thus demand years or decades of evolutionary time. These methods are ill-fit to identify early adaptive evolution of a virus that has experienced a recent spillover event, such as SARS-CoV-2, given that the common ancestor of globally circulating viruses corresponding to basal clade 20A or lineage B.1 is currently no earlier than January 2020 (`https://nextstrain.org/ncov/gisaid/global`). Here, we present a new method to identify regions of the genome undergoing adaptive evolution, which is well-suited to early time points. This method correlates clade success with the accumulation of protein-coding changes in certain genes. We apply this method to SARS-CoV-2 genomic data from Dec 2019 to May 2021, focusing on the period of VOC and VOI emergence.

We show that the association between clade growth rates and nonsynonymous mutations is highest within the S1 subunit, suggesting a positive fitness effect of S1 substitutions. Additionally, the ratio of nonsynonymous to synonymous divergence is markedly higher in S1 than other regions of the genome. We also examine the dynamics of adaptive evolution within the S1 subunit. Substitutions within S1 display a distinct pattern of temporal-clustering that synonymous mutations and RdRp substitutions do not. Several of these S1 substitutions, and a handful of mutations in other genes, exhibit convergent evolution, occurring

independently many times and giving rise to successful viral clades each time they do. One of these mutations is a 3-amino acid deletion in the Nsp6 gene (ORF1a:3675-3677del), which occurs at the base of over half of the VOC clades and precedes the accumulation of more S1 substitutions than almost any other convergently-evolved mutation. Together, these results indicate adaptation to a partially immune host population is sculpting the evolutionary trajectory of SARS-CoV-2.

4.2 Results

4.2.1 Accumulation of nonsynonymous mutations in spike S1 correlates with clade success

RNA viruses are known for their remarkably high error rates and, thus, the rapid generation of mutations. Despite possessing some proof-reading capacity (a relatively rare function for an RNA virus), SARS-CoV-2 has been accumulating roughly 24-25 substitutions per year (`https://nextstrain.org/ncov/gisaid/global?l=clock`). The null hypothesis is that these substitutions reflect neutral evolution: the result of genetic drift acting on random mutations. To determine whether this is true, or whether adaptive evolution is also contributing to the accumulation of mutations, we started by comparing substitution rates in different regions of the genome.

We built a time-resolved phylogeny with a balanced geographic and temporal distribution of isolates sampled between December 2019 and May 15, 2021 that includes 9544 viruses. Isolates are labeled by their emerging lineage membership- a designation which includes WHO VOCS, VOIs, and prominent PANGO lineages (Figure 4.S1). For every internal node on the phylogeny, we tallied the total number of mutations that occurred between the phylogeny root and that node. We grouped deletion events with nonsynonymous single nucleotide polymorphisms (SNPs), as they are protein-changing and contribute to the evolution of some regions of the genome (Figure 4.S2). Plotting mutation counts over time shows that spike S1 accumulates nonsynonymous changes at a rate of 8.4×10^{-3} substitutions/codon/year, or about 5.5 substitutions per year (Figure 4.1A). This is a disproportionate percentage of

the genome-wide estimate of 24 substitutions per year. As a control, we counted S1 synonymous mutations, and found they accumulate at 2.0×10^{-4} substitutions/codon/year, close to the naive expectation from base composition that 22% of mutations should be synonymous. The rate of nonsynonymous mutation in S1 is roughly 17 times higher than in the RNA-dependent RNA polymerase (RdRp) gene.

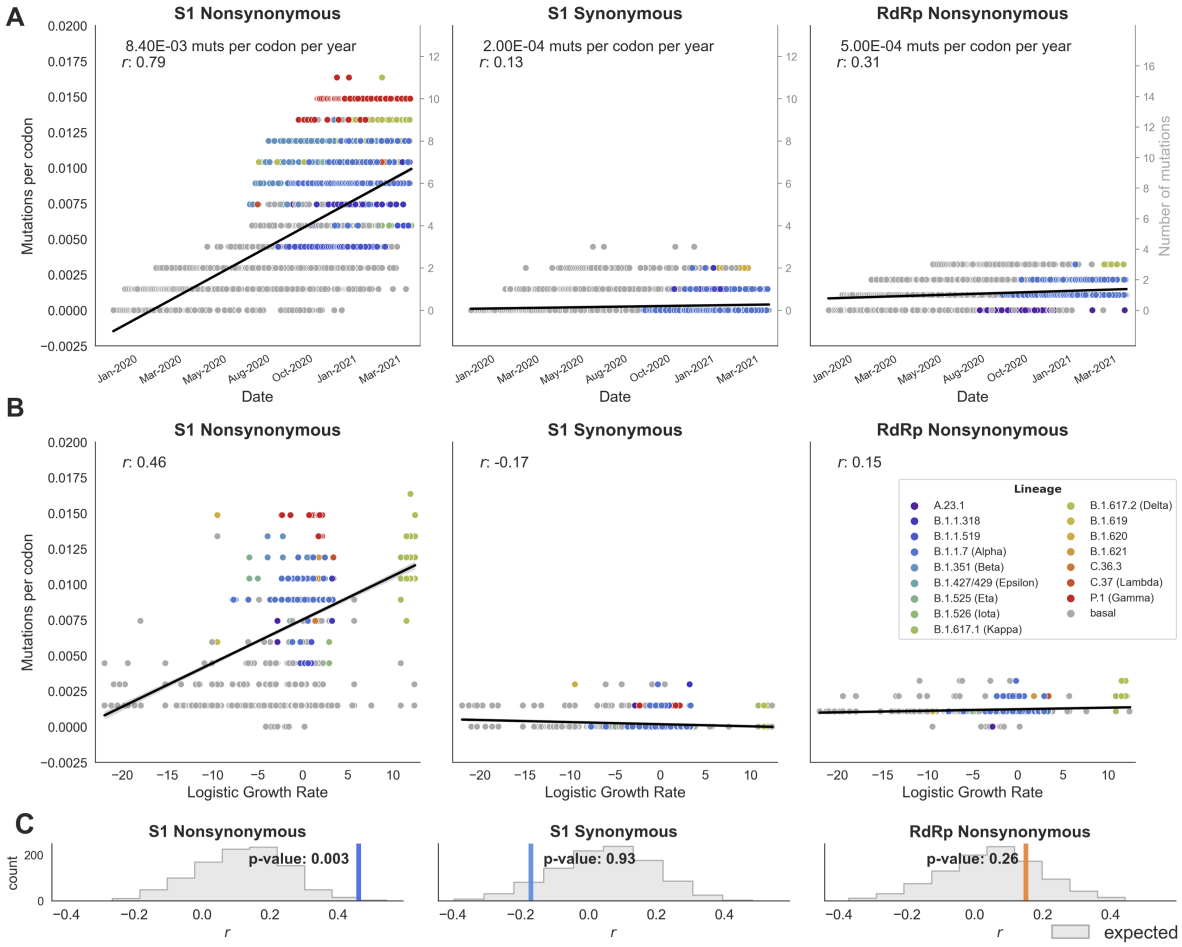

Figure 4.1: Accumulation of nonsynonymous S1 mutations is correlated with clade success. **A)** For every clade in the phylogeny, mutations relative to the root of the phylogeny are tallied and plotted against the date of the base of that clade. Nonsynonymous S1, synonymous S1, and nonsynonymous RdRp mutations are plotted separately. Nonsynonymous mutations include nonsynonymous SNPs and deletions. The primary axis (left, black ticks) displays mutations per codon, and the secondary axis (right, gray ticks) shows the absolute number of mutations accumulated in each clade. Each point is colored according to the lineage it belongs to. Points are fit by linear regression. **B)** For every clade, mutation accumulation (as in A) is plotted against logistic growth rate and the points are fit by linear regression. **C)**, The empirical correlation coefficient (r) between mutation accumulation and logistic growth rate (colored bar) is compared to an expected distribution (gray) to yield a p-value. Expected values of r are determined from randomizing mutations across the phylogeny using a multinomial draw with mutation likelihood proportional to relative branch length. The results of 1000 iterations are shown.

We hypothesize that adaptive evolution is driving the high rate of S1 nonsynonymous substitutions relative to S1 synonymous substitutions and RdRp nonsynonymous substitutions. If this is the case, we would expect a correlation between S1 substitutions and a clade's evolutionary success: clades that happened to accumulate more S1 substitutions should have, on average, higher fitness (and hence faster growth in frequency) than clades that have accumulated fewer S1 substitutions. Based on this logic, we introduce a new method for detecting adaptive evolution, which looks for regions of the genome where mutation accumulation is associated with clade frequency growth. Because positive selection causes alleles or clades to increase in frequency in a logistic (rather than linear) fashion, we measure logistic growth rate and plot this versus mutation accumulation.

Clade success and the number of nonsynonymous S1 mutations are positively correlated, with a correlation coefficient r of 0.46 (Figure 4.1B). To test whether this correlation is greater than expected, we randomized the positions of mutations across the phylogeny and computed a p-value between the empirical r and the distribution of r values from 1000 randomizations. The positive correlation between S1 mutations and logistic growth rate is statistically significant compared to the expected distribution ($p < 0.005$), but is absent for S1 synonymous mutations and is not significant for RdRp substitutions (Figure 4.1C).

We applied this method to every gene in the SARS-CoV-2 genome 4.1. The highest nonsynonymous mutation rate is observed in ORF8. However, ORF8 substitutions are not correlated with clade success 4.1, and many lineages acquire premature stop codons in ORF8, indicating that the high rate of ORF8 substitutions is likely due, at least in part, to a lack of functional constraints. Mutations within other regions of the genome, including spike S2 and nucleocapsid (N), also accumulate at reasonably-high levels but do not correlate well with clade success (Table 4.1). Besides S1, only Nsp6 (r=0.35, p=0.011) and ORF7a (r=0.43, p<0.001) have a strong correlation with clade growth rates (Table 4.1, and Figures 4.S3 and 4.S4).

Though ORF7a substitutions appear highly correlated with clade success, this correlation

Gene	Rate of nonsyn mutations (subs/codon/year)	r: nonsyn mut accumulation vs. logistic growth rate	p-value: empirical r versus expected
Nsp1	0.1×10^{-3}	0.05	0.431
Nsp2	0.2×10^{-3}	-0.1	0.881
Nsp3	1.3×10^{-3}	0.3	0.083
Nsp4	0.8×10^{-3}	0.28	0.057
Nsp5	0.5×10^{-3}	-0.09	0.833
Nsp6	3.4×10^{-3}	0.35	**0.011**
Nsp7	0.3×10^{-3}	-0.42	0.998
Nsp8	0.3×10^{-3}	0.04	0.443
Nsp9	0.7×10^{-3}	0.06	0.299
Nsp10	0.1×10^{-3}	-0.05	0.807
RdRp	0.5×10^{-3}	0.15	0.256
Nsp13	0.8×10^{-3}	0.2	0.153
Nsp14	0.3×10^{-3}	-0.03	0.698
Nsp15	0.3×10^{-3}	0.15	0.230
Nsp16	0.3×10^{-3}	-0.22	0.963
S1	8.4×10^{-3}	0.46	**0.003**
S2	3.5×10^{-3}	0.24	0.105
ORF3a	1.8×10^{-3}	-0.06	0.865
E	1.6×10^{-3}	0.03	0.388
M	0.7×10^{-3}	0.29	0.045
ORF6	0.2×10^{-3}	0.01	0.528
ORF7a	1.5×10^{-3}	0.43	**<0.001**
ORF7b	1.7×10^{-3}	0.22	0.050
ORF8	16.5×10^{-3}	0.19	0.185
N	6.10×10^{-3}	0.21	0.222
ORF9b	2.10×10^{-3}	0.27	0.050

Table 4.1: Genome-wide correlation between nonsynonymous mutation accumulation and logistic growth rate. For each gene (or subunit), the rate of nonsynonymous substitutions (and deletions) per codon per year is given. The correlation coefficient (r) of the linear regression between mutation accumulation and logistic growth rate and the p-value of this r compared to an expected distribution are listed.

is driven solely by the rapidly growing Delta variant, which possesses 3 mutations in ORF7a. Removing Delta clades from the analysis drops the r for ORF7a from 0.43 to 0.16, whereas r for S1 and Nsp6 only dip from 0.46 to 0.41, and from 0.35 to 0.32, respectively. This indicates that the correlation between S1 and Nsp6 substitutions and clade success is a general feature of SARS-CoV-2 lineages. Thus, the metric presented here provides evidence that SARS-CoV-2 is evolving adaptively and that the predominant locus of this evolution is spike S1.

4.2.2 The ratio of nonsynonymous to synonymous divergence is highest in S1

A classical method for assessing the directionality of natural selection on some region of the genome is d_N/d_S, measuring the divergence of nonsynonymous sites relative to synonymous sites. A d_N/d_S value less than 1 indicates that the region is, on average, under purifying

selection, while d_N/d_S greater than 1 indicates positive selection on the region. Because even the most rapidly evolving genes are still subject to structural and functional constraints, it is rare for an entire gene to have a d_N/d_S ratio greater than 1. For instance, the HA1 subunit of H3N2, which is the prototypical example of an adaptively-evolving viral protein, has d_N/d_S of 0.37 (Wolf et al., 2006).

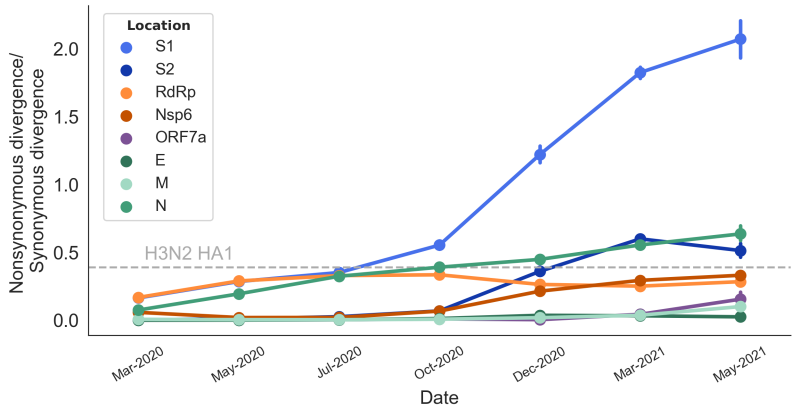

Figure 4.2: Ratio of nonsynonymous to synonymous divergence is highest S1. The ratio of nonsynonymous to synonymous divergence within various coding regions is plotted over time. Each point shows the mean and 95% confidence interval of this ratio for all internal branches present in a 0.2 year window (ending at the date indicated on the x-axis). Nonsynonymous divergence is calculated as the nonsynonymous Hamming distance between the sequence of an isolate and the root, normalized by the total possible number of nonsynonymous sites. The same is done for synonymous divergence. Divergence is calculated for various locations within the genome. The dashed line marks the average d_N/d_S of the influenza H3N2 HA1 subunit across all 12 years, as calculated in 4.S5.

For various regions of the SARS-CoV-2 genome, we computed the nonsynonymous to synonymous divergence ratios over the course of the pandemic thus far. The d_N/d_S ratio within RdRp, S2, and the structural proteins Envelope (E), Membrane (M), and Nucleocapsid (N) is consistently under 1 at all timepoints (Figure 4.2). However, d_N/d_S within S1 increases over time, with an apparent inflection point in mid-2020, and the d_N/d_S ratio exceeding 1 in late-2020 and 2021 with the most recent time point measured at 2.07. The increase over time in S1 d_N/d_S could be due to a variety of reasons. Two non-mutually exclusive hypotheses include the appearance of a new selective pressure on S1 substitutions, or the acquisition of mutations that change the mutational landscape to be more permissive towards S1 substitutions. Regardless of the cause, this change suggests a temporal structure

to the adaptive evolution in the S1 subunit of SARS-CoV-2.

4.2.3 Nonsynonymous mutations in spike S1 cluster temporally

A hint of this temporal structure can be seen by tracing individual mutational paths through the tree, from root to tip. Figure 4.S6 plots the accumulation of nonsynonymous S1 mutations along ten representative paths, leading to 10 different emerging lineages. Along each of these paths, there appears to be an initial period of relative quiescence, followed by a burst of S1 substitutions. To test whether this temporal clustering of mutations differs from what would be expected given the phylogenetic topology and the total number of observed S1 substitutions, we calculated wait times between mutations (diagrammed in Figure 4.3A). Briefly, we created a null expectation by running 1000 iterations of mutation randomization in which the phylogenetic location of every observed mutation is shuffled. The distribution of wait times is dependent on tree topology and total number of mutations, so the expectation is different for each category of mutations (Figure 4.S7).

If mutations are clustered, there should be an excess of short wait times in the empirical data relative to the expectation. This is what we observe for S1 nonsynonymous mutations, where the distribution of wait times is left-skewed, with an overabundance of short wait times compared to the expected distribution (Figure 4.3B). The mean wait time between observed S1 substitutions is significantly lower than the expected mean wait time ($p<0.001$), while there is no significant difference for S1 synonymous or RdRp wait times (Figure 4.3Ci). This difference is driven by short wait times because there is a significant difference between the proportion of observed versus expected wait times under 0.3 years for S1 nonsynonymous, but not S1 synonymous or RdRp, mutations (Figure 4.3Cii). These results indicate a temporal structure to the adaptive evolution of SARS-CoV-2 within the S1 subunit, which is characterized by mutation clustering.

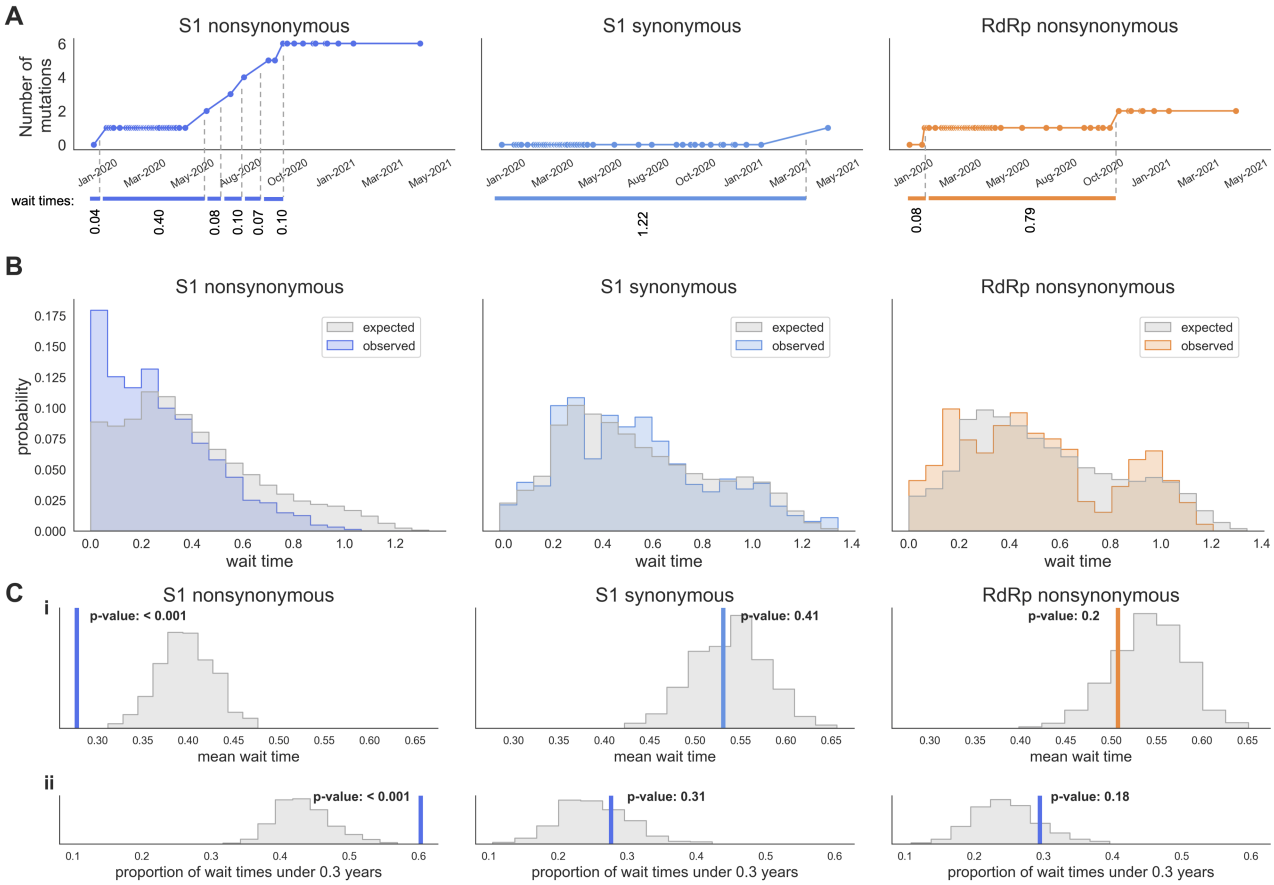

Figure 4.3: S1 substitutions are temporally clustered. A) Time line showing accumulation of S1 nonsynonymous, synonymous, and RdRp nonsynonymous mutations between the root and a representative tip (isolate USA/ME-HETL-J3202/2021) with wait times between mutations illustrated in below. The exact date of a mutation is chosen by randomly selecting a date along the branch the mutation occurs on. **B)** Distribution of wait times between S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous mutations. Empirical wait times (in color) are plotted along with expected wait times (gray). Expected wait times determined from randomizing mutations across the phylogeny using a multinomial draw with mutation likelihood proportional to relative branch length. The results of 1000 iterations are shown. **C) i)** The mean empirical wait time from 1000 iterations of the analysis (colored bar) is compared to the distribution of mean expected wait times (gray) to yield a *p*-value. **ii)** The proportion of observed wait times under 0.3 years (colored bar) is compared to the distribution of expected wait times under 0.3 years (gray).

4.2.4 Specific mutations associated with successful clades

We next sought to identify specific adaptive mutations throughout the genome. We note that convergent evolution is a good indicator of positive selection because each additional independent occurrence on the phylogeny of the mutation is increasingly unlikely under neutral evolution. As other groups have reported, there are many mutations shared by the VOCs that have arisen via convergent evolution (van Dorp et al., 2020; Rochman et al., 2021;

Martin et al., 2021). Here, we combine this observation of convergent evolution with logistic growth rate to find mutations that have arisen in the SARS-CoV-2 population multiple, independent times and expand into successful clades after each occurrence.

In this analysis, we focus on the evolutionary dynamics of SARS-CoV-2 during the period of time between the emergence of this virus in humans and mid-May 2021. We estimate that, during this period of time, VOC viruses are primarily competing with basal SARS-CoV-2 viruses. This allows us to examine the overall fitness effects of specific mutations in viral lineages that are successful during this period of time. After May 2021, VOCs comprise a majority of the global virus population, and similar analyses on later time points would speak to the relative competitiveness of the variants.

For every deletion and substitution observed on the phylogeny, we tallied the number of independent occurrences and found the mean logistic growth rate of all clades where this mutation occurred. We limited this analysis to internal branches with 15 or more descending samples to limit the influence of stochasticity and sequencing errors that often occur on terminal branches. As expected, the bulk (84%) of mutations occur just once. Roughly 4% of mutations arose 4 or more times, and the majority of these mutations are located in S1 (Figure 4.4A). For seven of these convergently-evolved mutations, the mean growth rate is higher than the tree-wide average growth rate. For three of these mutations (S:95I, S:452R and ORF1a:3675-3677del), the mean growth rate exceeds the 90th percentile of mean growth rates expected from a mutation that occurs the same number of times on a randomized tree (Figure 4.4B).

This analysis reveals influential mutations during a snapshot of time in the ongoing adaptive evolution of SARS-CoV-2. In mid-May 2021, the Delta variant was rising in frequency. Both S1 mutations we identified as important drivers of adaptive evolution (S:95I and S:452R) are present in the Delta variant as well as a handful of other emerging lineages (Figure 4.S8). The specific mutations identified by this analysis will vary over time and depend on a multitude of factors (genetic, epidemiological, and otherwise) that determine

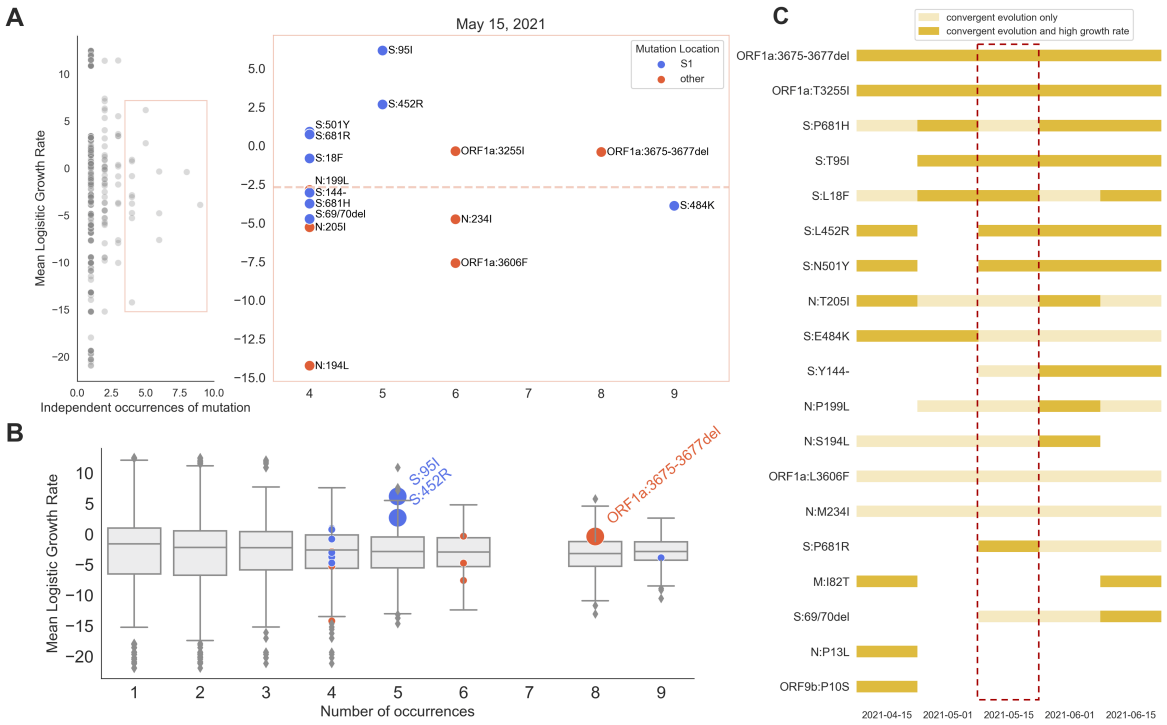

Figure 4.4: A 3-amino acid deletion in Nsp6 displays convergent evolution and occurs in successful clades. A) Every mutation observed on internal branches of the phylogeny is plotted according to the number of times this mutation occurs on the tree and the mean logistic growth rate of all clades it occurs in. Convergently-evolved mutations that appear 4 or more times across the phylogeny are shown in the inset. The average growth rate of all clades is shown with a dotted line. **B)** Observed mean growth rates of convergently-evolved mutations are compared to the mean growth rate expected for a mutation occurring on the phylogeny the same number of times. Convergently-evolved mutations that have a mean growth rate falling at or above the 90th percentile of the expected values are labeled. **C)** The analysis shown in panel A was completed for 5 time points, spanning two months. Each date represents the maximum date of sequences included in the analysis. Mutations that occur at least 4 times (convergent mutations) and result in a higher-than-average mean growth rate are shown in dark yellow. Mutations that display convergent evolution but do not result in high growth rates are in light yellow. The primary analysis was done using 2021-05-15 (outlined in red).

clade success. However, ORF1a:3675-3677del consistently appears as a top hit (Figure 4.4C, and Figure 4.S9). Remarkably, this deletion, which ablates amino acids 106-108 of Nsp6, arose 8 independent times and emerging lineages descend from each branch this deletion occurs on (Figure 4.S8).

Because recombination is common in coronaviruses (Müller et al., 2021; Turkahia et al., 2021), we investigated the possibility that these 8 occurrences of the ORF1a:3675-3677 deletion were due to recombination, rather than convergent evolution. We considered all pairs of lineages containing this mutation as potential recombinants and compared informative

mutations in the potential donor and acceptor. The closest informative mutations flanking ORF1a:3675-3677del are not shared by any pairs of lineages, offering a lack of evidence for recombination and strong support for convergent evolution.

4.2.5 A 3-amino acid deletion in nsp6 is associated with accumulation of S1 substitutions

The ORF1a:3675-3677 deletion in Nsp6 exhibits striking convergent evolution and consistently precedes successful viral lineages. Because we have shown that S1 mutation accumulation is also associated with clade success, we next asked whether there is a relationship between the number of S1 substitutions in clades containing ORF1a:3675-3677del.

We created an expectation for the mean number of S1 mutations that should be observed in clades with ORF1a:3675-3677del by generating 100 randomized trees where the mutation occurred on 8 branches selected by a multinomial draw. To make the expectation as fair as possible, we constrained the randomized branches to be on or after the date that the first Nsp6 deletion was observed. Under this expectation, there is no difference between the mean number of S1 or RdRp substitutions in clades that have the ORF1a:3675-3677 deletion versus clades that do not (Figure 4.5A, left). However, in the empirical phylogeny, there are significantly more S1 substitutions in clades with the Nsp6 deletion versus clades without (Figure 4.5A, right).

That clades with ORF1a:3675-3677del have higher numbers of S1 substitutions does not speak to the directionality of this relationship. In other words, it is possible that ORF1a:3675-3677del occurs in lineages that already have a lot of S1 substitutions, or that a lot of S1 mutations accumulate in clades that already have ORF1a:3675-3677del. To determine the directionality of this difference, we considered every phylogenetic path that contains the Nsp6 deletion and found the difference between the final number of S1 substitutions on that path and the number of S1 substitutions that had accumulated before the deletion. On average, around 2.5 S1 nonsynonymous mutations accumulate after ORF1a:3675-3677del (Figure 4.5B). This is the second largest increase in S1 mutation accumulation following

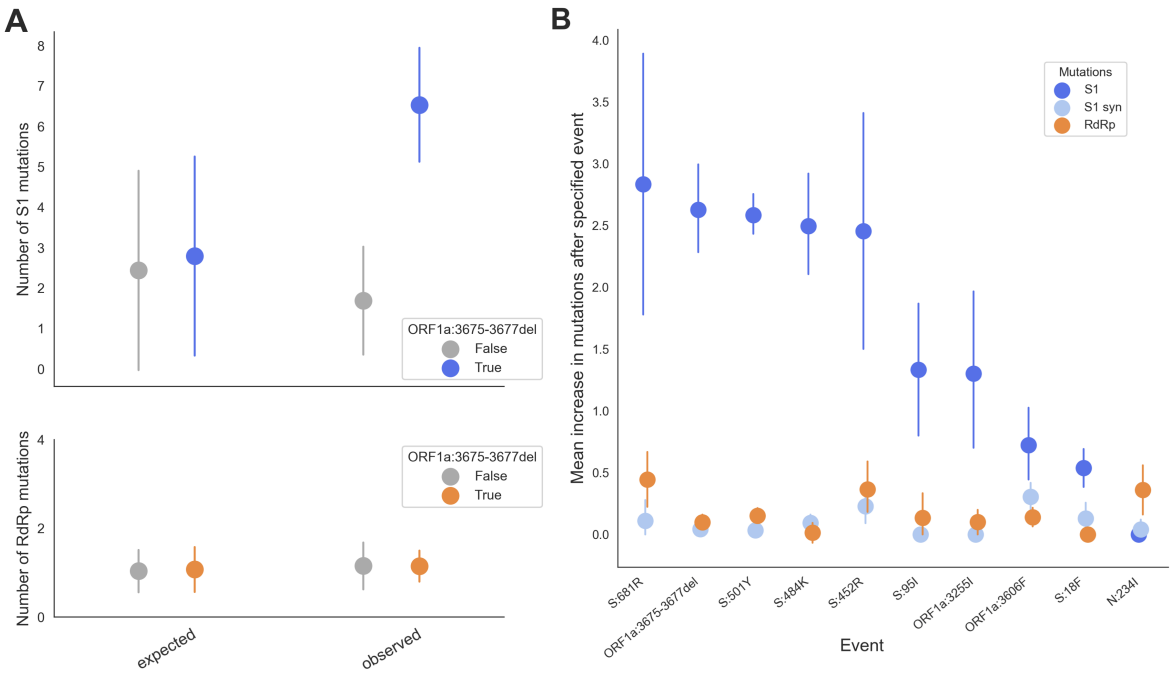

Figure 4.5: Clades with the 3-amino acid deletion in Nsp6 have a high number of S1 mutations. **A)** The mean number of S1 mutations (top), or RdRp mutations (bottom), that occur in clades that have (blue/orange) or do not have (gray) the 3-amino acid deletion in Nsp6. The expected difference is shown on the left, and empirical data is shown on the right. Expectation is based off of 100 randomizations of the tree. Error bars shown standard deviation. **B)** The difference in the number of nonsynonymous S1 (dark blue), S1 synonymous (light blue), and nonsynonymous RdRp (orange) mutations that occur before versus after a convergently-evolved mutation is shown. Error bars show 95% confidence intervals.

any convergently-evolved mutation, behind S:681R. These results do not indicate that the deletion directly causes S1 substitutions, but they do add to the observations of convergent evolution and high clade growth rates in suggesting that ORF1a:3675-3677del is an adaptive mutation and an influential factor in the evolution of SARS-CoV-2.

4.3 Discussion

Detecting adaptive evolution is both highly interesting from a basic scientific perspective as we seek to understand how and when this type of evolution occurs, and highly relevant from a public health perspective as we strive to curb the transmission of infectious diseases. As the SARS-CoV-2 pandemic rages on, our best defense is through vaccination. The SAR-CoV-2 vaccines showed high efficacy in clinical trials, but we must be proactive to ensure their continued effectiveness. Vaccines against viruses that undergo adaptive evolution at

antigenic sites, like influenza, must be continually updated to match circulating variants.

SARS-CoV-2 exhibits convergent evolution (van Dorp et al., 2020; Martin et al., 2021; Rochman et al., 2021), and some of the notable mutations that have occurred multiple times independently (like S:501Y and S:484K) appear in multiple VOCs, suggesting positive selection on these mutations. In the context of deep mutational scanning (DMS) experiments, mutations at 501 increase ACE2 binding affinity (Starr et al., 2020) and mutation to site 484 escapes antibody binding (Greaney et al., 2021). Recurrent mutations at S:681 enhance S1/S2 subunit cleavage (Lubinski et al., 2021; Liu et al., 2021), a protein-modification that is essential for spike-mediated cell entry (Hoffmann et al., 2020) and thus is thought to contribute to increased viral replication (Liu et al., 2021). Many other convergently-evolved mutations are also shared by VOCs and possess demonstrably different phenotypes, often altering antigenicity (Li et al., 2020; McCarthy et al., 2021; Wang et al., 2021).

Despite the demonstrably advantageous effects of observed mutations, it is too soon, evolutionarily, to pick up strong signals of adaptive evolution by the classical methods that rely on fixation of nonsynonymous mutations. Instead, we capitalize on the high temporal and geographic density of SARS-CoV-2 sequencing data to create a new method for identifying adaptive evolution and regions of the genome where this evolution is localized. This method identifies genes where amino acid substitutions significantly correlate with clade growth rate. This can be intuitively interpreted as genes with high rates of amino acid substitutions (suggestive of positive selection) that result in more successful viruses (suggestive of a positive fitness effect) are undergoing adaptive evolution. We find that the spike S1 subunit shows strong signals of adaptive evolution by this method (Figure 4.1).

Interestingly, we find temporal structure to this adaptive evolution. Substitutions within S1 cluster temporally (Figure 4.3), rather than accruing at a steady rate. The ratio of nonsynonymous to synonymous divergence (d_N/d_S) in S1 also increases over time (Figure 4.2). This temporal structure likely indicates a changing evolutionary landscape: either through the emergence of new selective pressure, and/or through the occurrence of permissive mu-

tations that made adaptive mutations more accessible. Additionally, selective pressure may be heterogeneous across the SARS-CoV-2 phylogeny due to particular transmission chains transiting through populations with greater seroprevalence. Our results do not distinguish between these possibilities.

While the overall d_N/d_S ratio in S1 is 0.76, d_N/d_S is 1.85 in 2021 (Figure 2). This high ratio is remarkable when compared to the antigenically-evolving HA1 subunit of influenza H3N2. We estimate the d_N/d_S ratio for HA1 to be 0.39 (Figure 4.S5), which is similar to the 0.37 estimated previously (Wolf et al., 2006). However, influenza H3N2 has been endemic in the human population for over 50 years, and it's current evolution is largely driven by antigenic changes (Smith et al., 2004). It is unclear whether this high d_N/d_S ratio in SARS-CoV-2 S1 will persist or whether it is a feature of this virus's recent emergence and will drop in the months and years to come.

An initially high rate of protein-coding changes is consistent with the idea that, soon after a spillover event, there are many evolutionarily-accessible mutations that are advantageous in the new host environment. This was observed in the influenza H1N1 pandemic virus (H1N1pdm). For 2 years following its emergence in 2009, H1N1pdm had elevated genome-wide d_N/d_S rates, and evolution during this period is thought to largely have been adaptation to a new host, including increased transmission in humans (Su et al., 2015). From 2011 onward, the adaptive evolution of H1N1pdm has been dominated by antigenic changes (Su et al., 2015). It is possible that SARS-CoV-2 is following a similar trajectory of adaptive evolution, with initial host adaptation to be followed by sustained antigenic drift.

Together, the results presented in Figures 4.1-4.3 offer phylogenetic evidence that SARS-CoV-2 is evolving adaptively and that the primary locus of this adaptation is in S1. This is consistent with experimental demonstration of phenotypic changes conferred by VOC spike mutations (Wang et al., 2021; Greaney et al., 2021; Li et al., 2020; Liu et al., 2021). Adaptive evolution in the S1 subunit is likely driven by selection to increase cell infectivity, and/or to escape neutralizing antibodies. These functions are not mutually exclusive, and it has

been shown that selection for binding affinity in H3N2 yields mutations that incidentally evade humoral immune recognition (Hensley et al., 2009). The potential antigenic impact of adaptive S1 mutations, which are accruing at pace over 4 times that of influenza H3N2 (Figure 4.2, Figure 4.S5), suggests that it may be necessary to update the SARS-CoV-2 vaccine strain.

In addition to S1, our results suggest that substitutions within Nsp6 and ORF7a may significantly contribute to the success of viral clades (Table 4.1). We expand on these gene-wide results by identifying specific adaptive mutations, using the confluence of convergent evolution and clade success. This analysis turned up many S1 mutations that have been extensively studied, along with mutations to nucleocapsid (N), another target of antibody-recognition (Kang et al., 2021), and a couple mutations in Nsp6, Nsp4 and M (Figure 4.4). The non-S1 mutations ORF1a:3255I (in Nsp4), M:82T, and N:205I in particular show compelling evidence of positive selection. These sites enrich our understanding from gene-wide analyses presented in Figures 4.1-4.3 and Table 4.1: though S1 is the primary genomic locus of adaptive evolution, a handful of positively-selected mutations in other genes are also influencing the evolution of SARS-CoV-2 in the human population.

Our analysis of specific adaptive mutations suggests the possibility of differences between within-host selection for viral replication and between-host selection for transmission. Viruses belonging to Delta have shown greater between-host transmission rates than other VOC or VOI viruses (Campbell et al., 2021), but are lacking mutations that have occurred repeatedly and that were associated with increased clade growth (notably ORF1a:3675-3677del, S:484K and S:501Y). It is possible that some mutations display a large degree of parallelism due to specific within-host pressures that occur in secondary infections of partially immune individuals, despite having only modest effects on between-host transmission.

It is important to note that the precise mutations that appear most influential depend on when the analysis is done (Figure 4.4C and Figure 4.S9). The fitness effect of a mutation is not an absolute quality- it depends on a multitude of influences including genetic

background of the viral lineage, other co-circulating lineages, existing host immunity, and epidemiological factors (such as geographically heterogeneous mitigation efforts). Additionally, lineages can grow in frequency due to stochastic effects. It is, therefore, expected that mutations associated with successful clades will change over time and that these changes reflect both a changing fitness landscape and the stochastic nature of evolution. Mutations that transcend this or, in other words, are associated with successful lineages at multiple time points, are more likely to have important, adaptive functions. One such mutation is ORF1a:3675-3677del (Figure 4.4C and Figure 4.S9).

The ORF1a:3675-3677 deletion removes 3 amino acids (SGF) from a predicted transmembrane loop (Benvenuto et al., 2020) of the Nsp6 protein. Across the coronavirus family, the Nsp6 protein, in coordination with Nsp3 and Nsp4, forms double-membrane vesicles that are sites for viral RNA synthesis (Snijder et al., 2020). In SARS-CoV-2, Nsp6 suppresses the interferon-I response (Xia et al., 2020). It is unclear whether ORF1a:3675-3677del impacts either of these functions.

This deletion is not observed in other sarbecoviruses, residues 3675 and 3676 are 100% conserved, and only synonymous and conservative changes are seen at 3677 in this subgenus (Jungreis et al., 2021). However, in SARS-CoV-2, this deletion exhibits close to the highest level of convergence, presence in VOCs, mean logistic growth rate, and increase in S1 mutations in descending lineages. So far, ORF1a:3675-3677del has not been observed in Delta viruses and our results suggest that the appearance of a sublineage of Delta possessing ORF1a:3675-3677del may outcompete basal Delta viruses. Future experimental study of this deletion would increase our understanding of what functions, apart from enhanced cell entry and potential antibody escape, were highly advantageous during the early adaptive evolution of SARS-CoV-2.

4.4 Methods

The code for all analyses presented in this manuscript is located at `https://github.com/blab/sarscov2-adaptive-evolution`.

4.4.1 Phylogenetic reconstruction of a subsampling of global SARS-CoV-2 genome sequences

All analyses in this manuscript were performed using data downloaded from the GISAID EpiCoV database (`https://gisaid.org`, (Shu and McCauley, 2017)) on July 29, 2021 and curated by the Nextstrain nCoV ingest pipeline (`https://github.com/nextstrain/ncov-ingest`). This dataset contained 2,459,376 viral genomes and associated metadata. These genomes were aligned with Nextalign (`https://docs.nextstrain.org/projects/nextclade/en/latest/user/nextalign-cli.html`) and masked to minimize error in phylogenetic inference associated with problematic amplicon sites. Masked alignments were filtered to exclude strains that were known outliers, sequenced due to 'S dropout', misannotated with a admin division of 'USA', shorter than 27,000 bp of A, C, T, or G bases, missing complete date information, annotated with a date prior to October 2019, flagged with more than 20 mutations above the expected number based on the mutational clock rate, or flagged by Nextclade (`https://docs.nextstrain.org/projects/nextclade/en/latest/user/algorithm/07-quality-control.html`) with one or more clusters of 6 or more private differences in a 100-nucleotide window. After filtering 2,213,085 genomes remained.

After filtering, SARS-CoV-2 genomes were evenly sampled across geographic scales and time. Specifically, a maximum of 1,600 strains were sampled from each continental region including Africa, Asia, Europe, North America, Oceania, and South America for an approximate total of 9,600 genomes per phylogeny. For each region except North America and Oceania, strains were sampled from each distinct combination of country, year, and month. For North America and Oceania, genomes were sampled from each distinct combination of division (i.e., state-level geography), year, and month.

Time-resolved phylogenies were inferred using Augur 12.0.0 (Huddleston et al., 2021), IQ-

TREE 2.1.2 (Nguyen et al., 2015), and TreeTime 0.8.2 (Sagulenko et al., 2018). Ancestral sequences were inferred with TreeTime using the joint inference mode. The primary analysis was conducted on 9544 genomes collected on or before May 15, 2021, and the phylogeny reconstructed from these data can be found at `https://nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15`. Phylogenies used for secondary analyses of convergent evolution (Figure 4.4C, and Figure 4.S9) can be viewed using the date drop-down menu in the left-hand sidebar. The secondary analyses included isolates sequenced up until April 15, 2021 (9467 genomes), May 1, 2021 (9449 genomes), June 1, 2021 (9343 genomes), and June 15, 2021 (9401 genomes). All isolates used in these analyses are listed in the Acknowledgements table found at `https://github.com/blab/sarscov2-adaptive-evolution/blob/master/sars2_manuscript/sars2_adaptive-evolution_acknowledgements.tsv`.

Influenza H3N2 trees (used for Figure 4.S5) were run by cloning the `https://github.com/nextstrain/seasonal-flu/` repo and running builds for HA1 and PB1 with 12 year resolution.

4.4.2 Quantification of mutation accumulation

For every internal branch on the phylogeny, the number of mutations that accumulated between the root of the tree and that branch was counted. For this and all subsequent analyses, deletions are grouped with nonsynonymous substitutions. Deletions that span multiple, adjacent amino acids are counted as one mutation. Mutations to a premature stop codon are also counted as one mutation event. Mutations were separated by which gene they occur in (according to the Wuhan-Hu-1 reference sequence, found at `analysis/reference_seq_edited.gb`) and whether they are synonymous or nonsynonymous. Genomic locations of the 15 NSPs were found in the NC_045512.2 annotation of the ORF1ab polyprotein (`https://www.ncbi.nlm.nih.gov/gene/43740578`). Code for mutation accumulation counting and plotting of Figure 4.1A is found in `fig1-muts_by_time_and_growthrate`.

4.4.3 Estimation of the logistic growth rate of clades

Logistic growth of individual clades was estimated from the time-resolved phylogeny and the estimated frequencies for each strain in the tree. Frequencies were estimated with Augur 12.0.0 (Huddleston et al., 2021) using the KDE estimation method that creates a Gaussian distribution for each strain with a mean equal to the strain's collection date and a variance of 0.05 years. At weekly intervals, the frequencies of each strain at a given date were calculated by summing the corresponding values in their Gaussian distributions and normalizing the values to sum to 1. The frequency of each clade at a given time was the sum of its corresponding strain frequencies at that time.

Logistic growth was calculated for each clade in the phylogeny that was currently circulating at a frequency >0.0001% and <95% and that had at least 50 descendant strains. Each clade's frequencies for the last six weeks were logit transformed and used as the dependent variable for a linear regression where the independent variable was the corresponding date value for each transformed frequency. The logistic growth of the clade was then annotated as the slope of the linear regression of the logit-transformed frequencies.

4.4.4 Calculation of nonsynonymous to synonymous divergence ratio

A time-course of d_N/d_S ratios was calculated in non-overlapping time windows by splitting all internal branches (with 3 or more descending tips) included in the phylogeny according to their date. Within each gene, the nonsynonymous and synonymous Hamming distances were found between the reference sequence and every internal branch. The Hamming distances were normalized by the total number of possible nonsynonymous or synonymous sites within that gene to give a measure of divergence. The nonsynonymous divergence was divided by synonymous divergence. Then, for each time window, the mean of this ratio was found for all internal branches within the window. For SARS-CoV-2, the time windows were 0.2 years and the code to run this analysis and reproduce Figure 4.2 is at `fig2-divergence.ipynb`. For H3N2, the time windows were 0.4 years and the code is in

`fig2supp-divergence_h3n2.ipynb`.

4.4.5 Randomization of mutations across the phylogeny for wait time calculations

For each type of mutation (S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous), the total number of mutations observed on the phylogeny was randomly scattered across phylogeny. Only internal branches with 3 or more descending tips were used. Random branches were selected by a multinomial draw, where the likelihood of a branch having a mutation is proportional to its branch length in years. Multiple mutations were allowed to occur on the same branch, just as with the empirical phylogeny. Randomizations were run 1000 times for each mutation type used in Figure 4.3B and C, and 10 times for the distributions shown in figure 4.S7. Code for this analysis is in `fig3-wait_times.ipynb`.

4.4.6 Calculation of wait times

Wait times were counted for the following classes of mutations: S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous. For each class of mutation, a wait time was calculated between each branch that has a mutation of this type and its first child branch on each descending path that has a mutation of this type. A wait time was also calculated between the tree root and the first branch on any independent path that has a mutation of this type. Conceptually, the result of this is that wait times are computed between every sequential mutation that occurs along every path on the tree (as diagrammed in Figure 4.3A), without double counting any pairs of branches. Only mutations on internal branches (defined as having 3 or more descending tips) are considered.

A wait time is simply the time between mutations and is calculated by subtracting the date (in decimal years) of the earlier mutation from the date of the later mutation. Because the exact date a mutation occurred cannot be known, each mutation is assigned a random date along the branch it occurred on. If multiple mutations of the same type occurred on one branch, each mutation is assigned a different random date and the wait times between

mutations on that branch are calculated.

Empirical and expected wait times were calculated for each type of mutation 1000 times and the results of all 1000 iterations can be found in `wait_time_stats/`. Code to calculate wait times and reproduce Figure 4.3B and C and Figure 4.S7 is found in `fig3-wait_times.ipynb`.

4.4.7 Quantification of convergent evolution and logistic growth rates across the phylogeny

Every substitution that occurred on an internal branch with at least 15 descending tips was tallied. For every substitution that was observed at least 4 times on internal branches, the average growth rate of clades containing this mutation was calculated by taking the mean logistic growth rate of clades where this mutation occurred. Code to count occurrences, calculate mean logistic growth, and determine which emerging lineages descend from recurrent mutations is found in `fig4-convergent_evolution.ipynb`. This code will reproduce Figures 4.4A, 4.S8, and 4.S9.

4.4.8 Randomization of recurrent mutations across the phylogeny

One hundred randomized trees were created by shuffling the phylogenetic positions of each substitution that was observed on an internal branch with at least 15 descending tips (those calculated above and shown in Figure 4.4A). Randomized branches were also limited to internal branches with at least 15 descending tips. The position of each randomized substitution was constrained to branches that "make phylogenetic sense": meaning, a given substitution cannot occur twice on the same path. This results in a tree with exactly the same distribution of mutation occurrences as the empirical phylogeny, but where those mutations occur on different branches. Code to implement these randomizations and reproduce Figure 4.4B is in `fig4-convergent_evolution.ipynb`.

4.4.9 Consideration of recombination as an alternative to convergent evolution of nsp6 deletion

For each occurrence of the ORF1a:3675-3677 deletion, all nucleotide mutations that occurred between the root and the branch where the deletion occurred were recorded. Then, recombination between every pair of the 8 inferred occurrences of ORF1a:3675-3677del was considered. For each pair, informative mutations that did not occur in a common ancestor of the potential recombinant lineages were identified. The informative mutations closest to the Nsp6 deletion on the upstream side were compared between potential donor and acceptor (and the same was done for the downstream side). If the closest mutations were shared between any donor/acceptor pair, this would be evidence that this mutation and the Nsp6 deletion were transferred from the donor to the acceptor by recombination. If the closest mutations are not shared between the donor and acceptor, the only way the acceptor could have acquired the ORF1a:3675-3677del through recombination is if both recombination break points occurred within a genomic window defined by the closest informative mutations on either side of the Nsp6 deletion. Code for this analysis as well as a table summarizing the results is in `nsp6del_recombination.ipynb`.

4.4.10 Calculation of the mean number of S1 mutations per clade

The phylogeny was divided into clades that have the ORF1a:3675-3677 deletion and those that do not, and the mean number of S1 and RdRp substitutions was computed for each category. The tree was limited to only branches occurring on or after the date of the first ORF1a:3675-3677del occurrence. The expectation was created by randomizing the locations of the 8 occurrences of ORF1a:3675-3677del as was done above in "Randomization of recurrent mutations across the phylogeny". Code for this analysis is in `fig5a-nsp6del_s1mutations_correlation.ipynb`.

4.4.11 Calculation of S1 mutations that precede and follow specific mutation events

For each convergently-evolved mutation, every path through the phylogeny containing this mutation was considered. The total number of S1 mutations accumulated between the root and the occurrence of the convergently-evolved mutation is considered to be the number of S1 mutations before the event. The number of mutations after is the final number of S1 mutations present on the path. The before total is subtracted from the after total to give the increase in S1 mutations after the event. The mean of this increase is calculated for every path containing the convergently-evolved mutation. Code to implement this analysis is in `fig5b-s1_muts_before_vs_after.ipynb`.

4.5 Supplemental Information

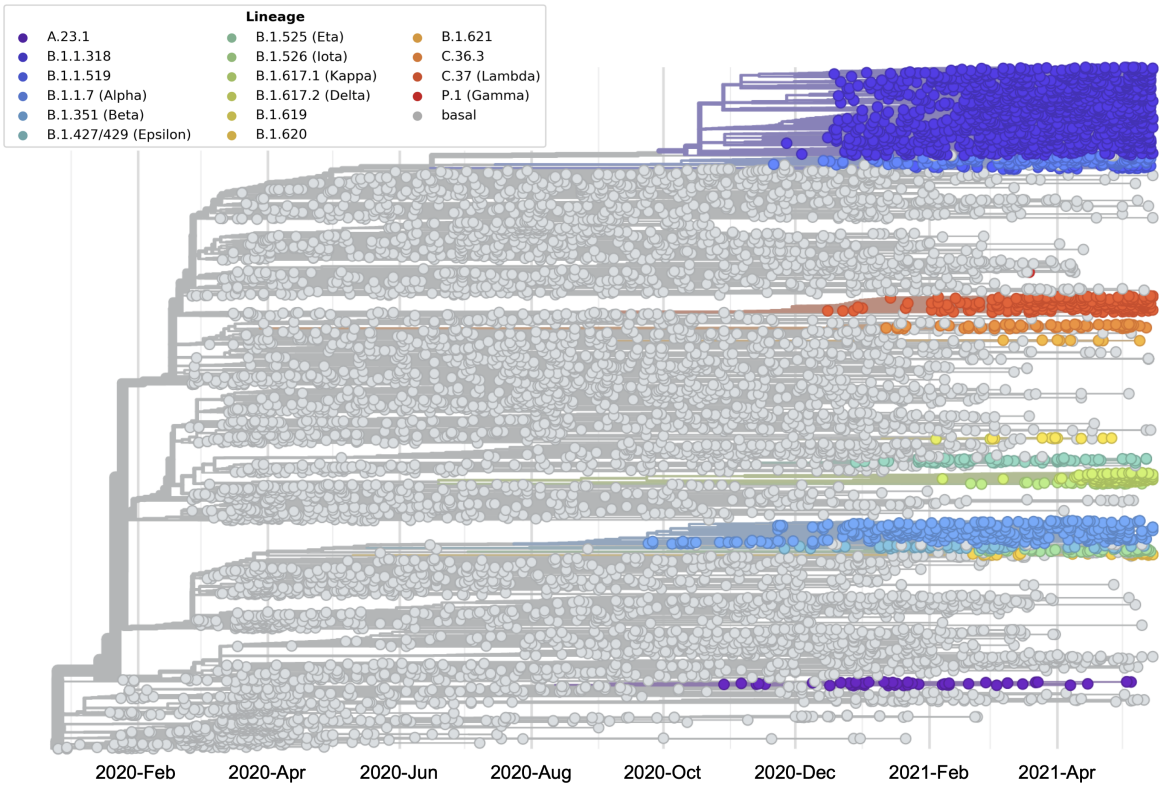

Figure 4.S1: Phylogeny of 9544 SARS-CoV-2 genomes. Screenshot of the phylogeny used for the primary analyses in this manuscript. Tips and branches are colored according to emerging lineage. Emerging lineages are labeled by PANGO lineage and WHO Variant of Interest (VOI) or Variant of Concern (VOC) designation. An interactive version of this phylogeny can be accessed at nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15.

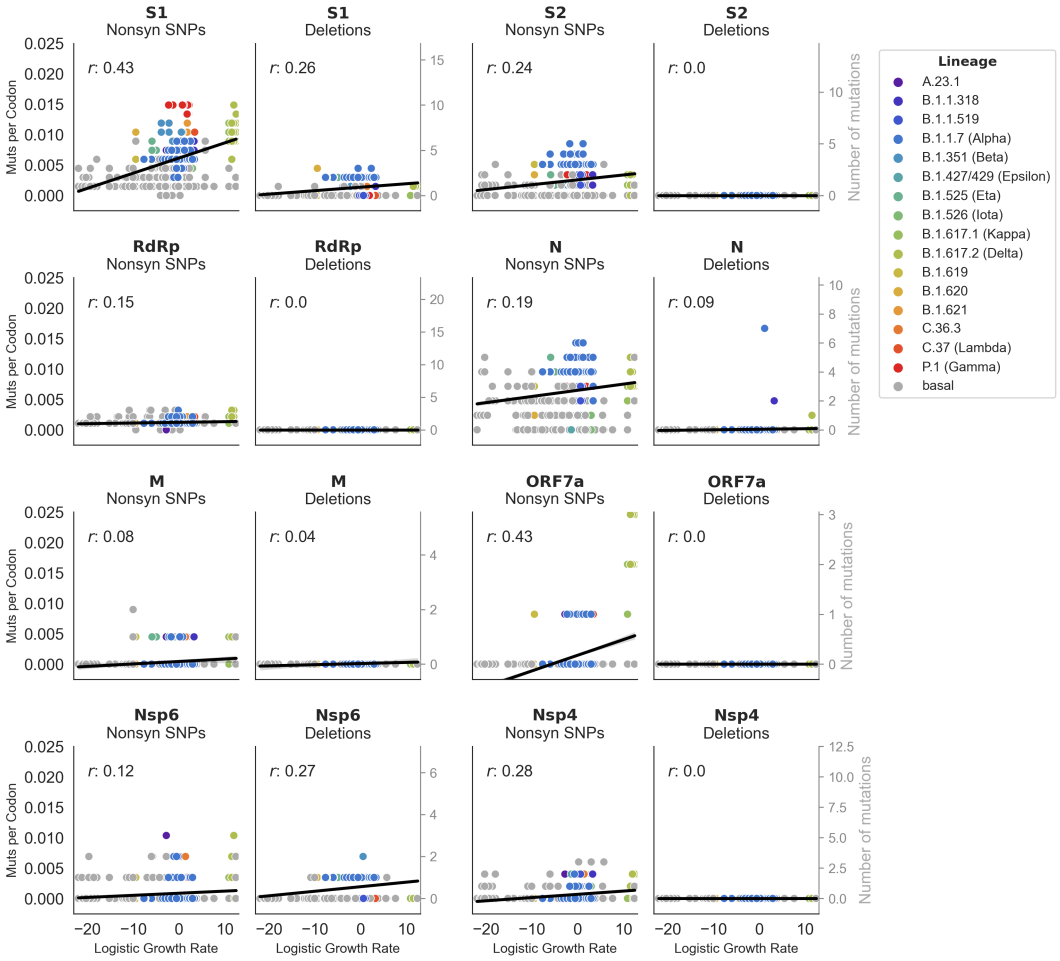

Figure 4.S2: Deletions contribute to protein-coding changes in S1, N and Nsp6 For each gene, nonsynonymous mutation accumulation is separated into nonsynonymous SNPs (left) and deletions (right). Accumulation of these mutations is plotted against logistic growth rate for 8 genes (or subunits), as in Figure 4.1B.

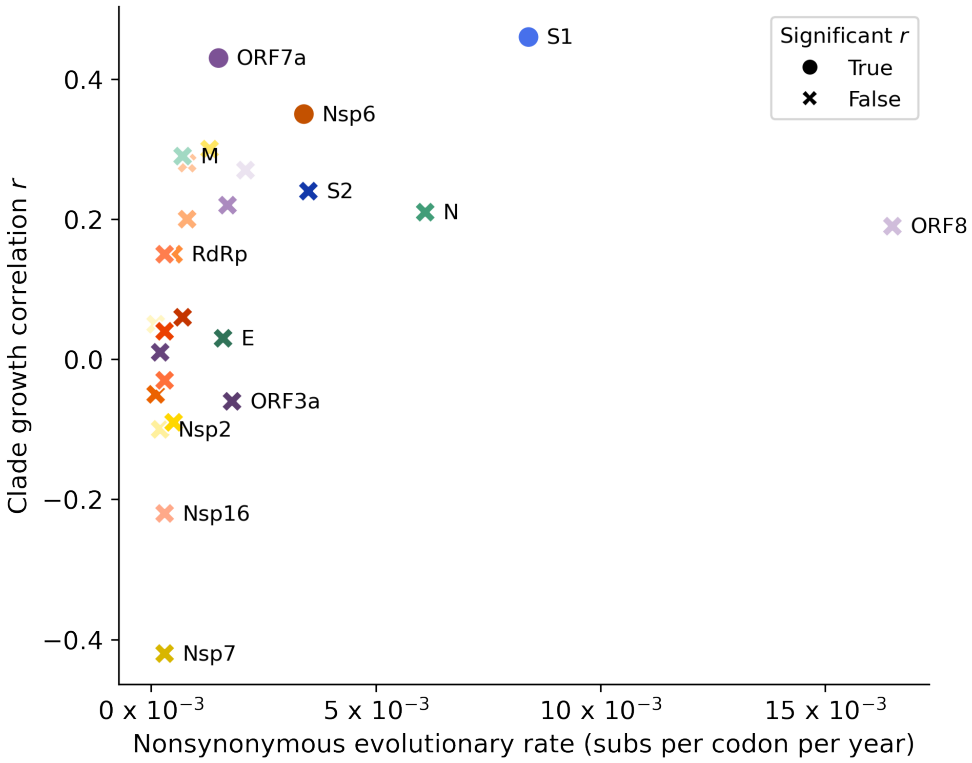

Figure 4.S3: Visual Representation of Table 4.1. For every gene in the genome, the rate of nonsynonymous substitutions (and deletions) per codon per year is plotted against the correlation coefficient r of mutation accumulation with logistic growth. Circles indicate genes with significant r values at the p=0.01 level, and Xs indicate genes with insignificant r values.

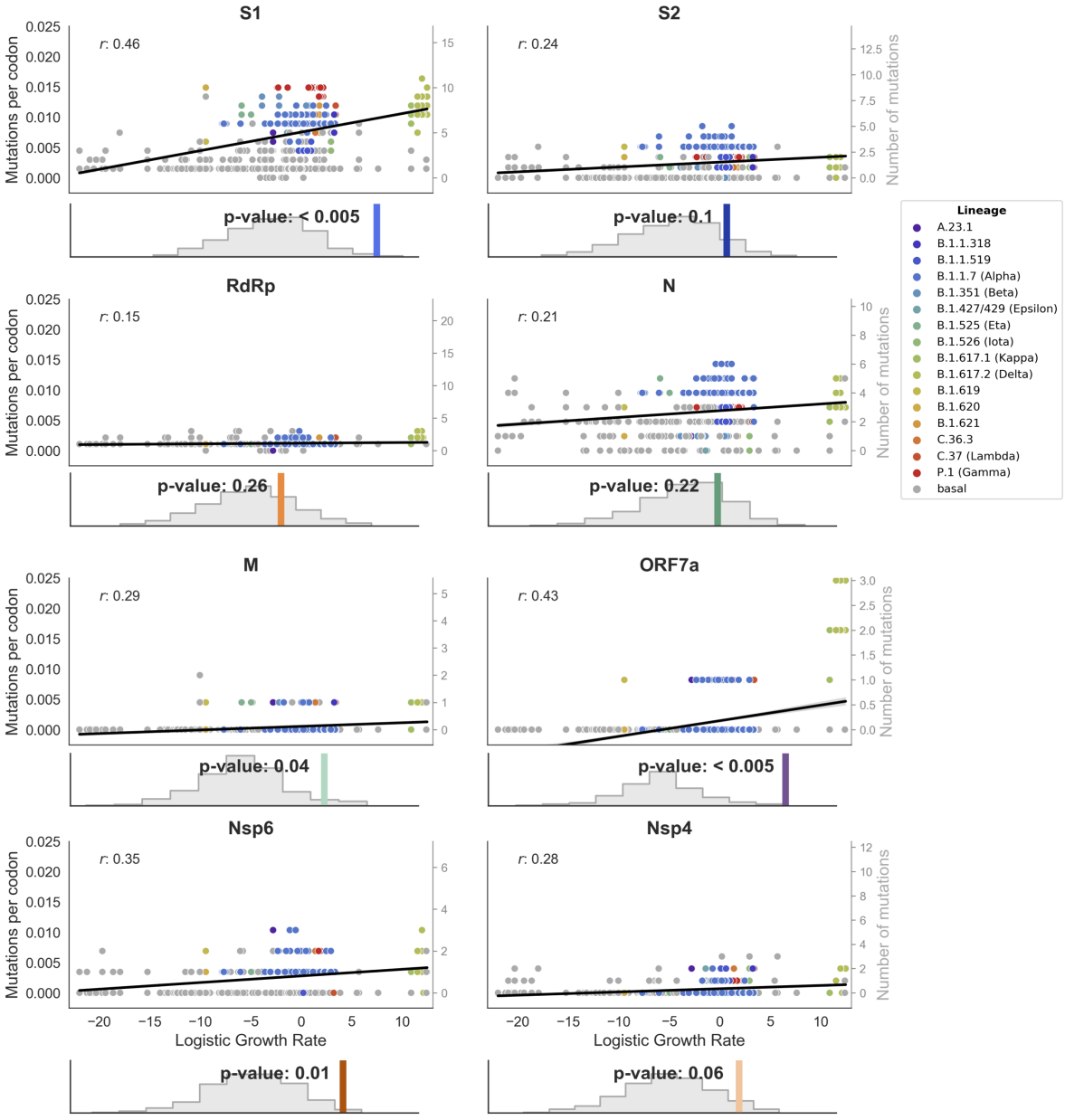

Figure 4.S4: Correlation between nonsynonymous mutation accumulation and clade success is strongest in S1. Nonsynonymous mutation accumulation (mutations per codon) is plotted against logistic growth rate for 8 genes (or subunits), as in Figure 4.1B. Histograms beneath each plot show the empirical r-value (colored line) compared to the distribution of r-values from 1000 randomizations, as well as the *p*-value resulting from this comparison.

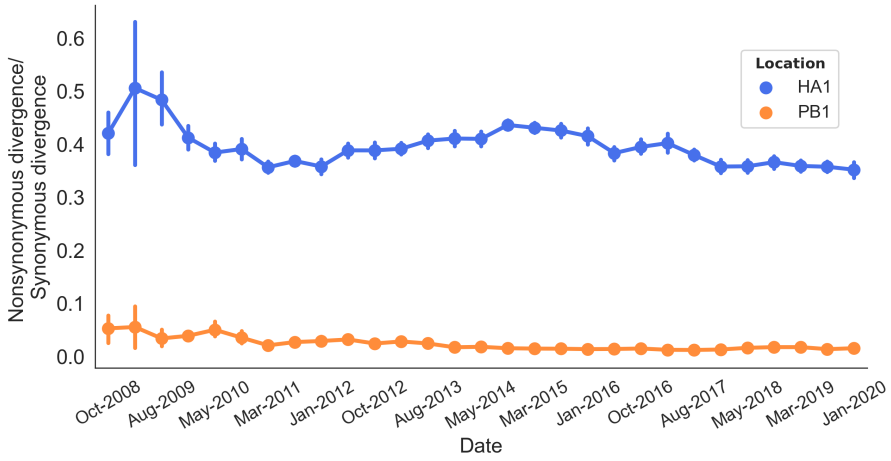

Figure 4.S5: Ratio of nonsynonymous to synonymous divergence in influenza H3N2. The mean and 95% confidence intervals for nonsynonymous/synonymous divergence ratios in the H3N2 genes HA1 and PB1 are shown over a 12-year period.

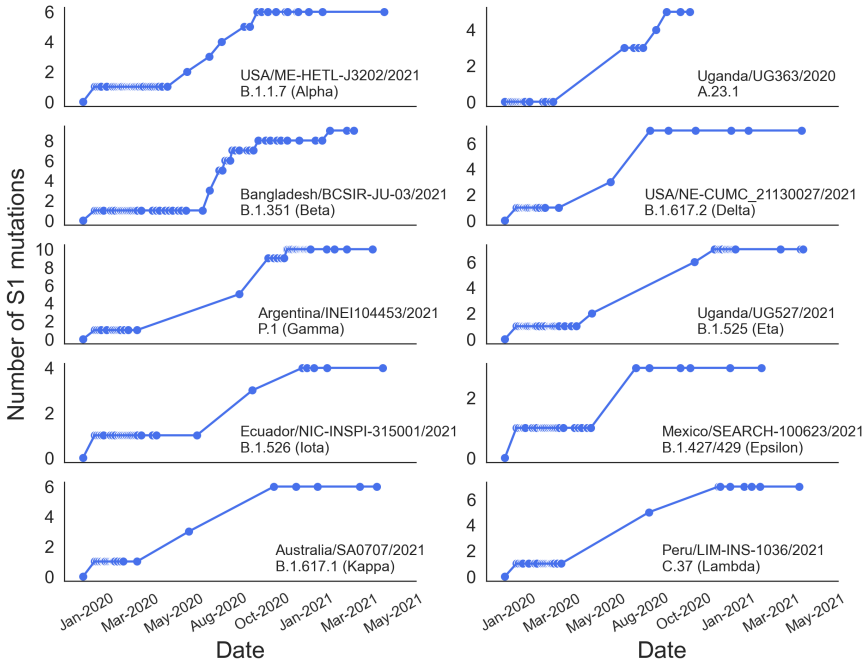

Figure 4.S6: Temporal accumulation of S1 mutations on representative paths through the tree. The total number of accumulated S1 nonsynonymous mutations is counted at every branch along a path through the tree. This is plotted for 10 representative paths from the root to an isolate in an emerging lineage clade. The isolate and emerging lineage are labeled on each panel.

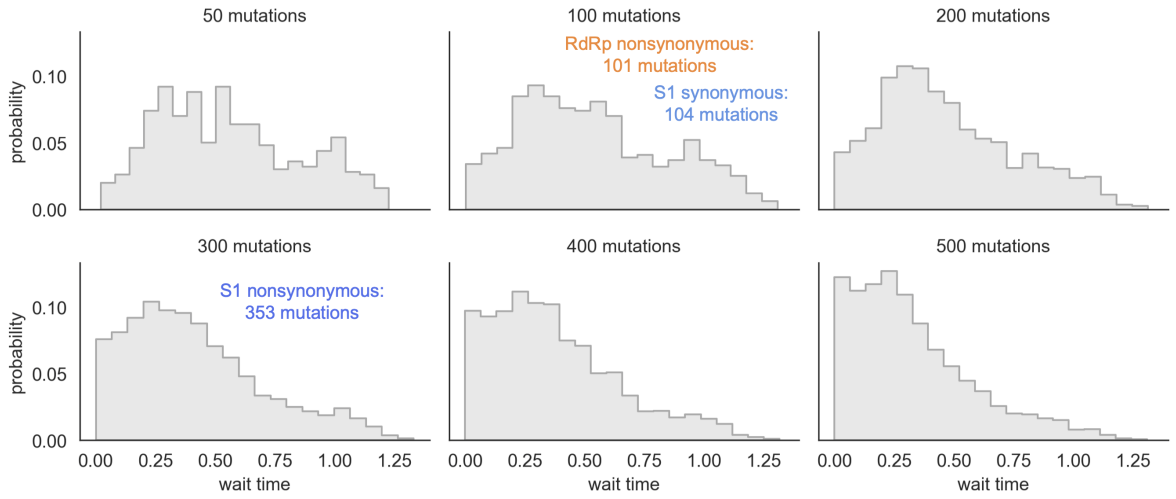

Figure 4.S7: Distribution of expected wait times is affected by the number of mutations that occur across the phylogeny. The phylogeny was randomized with varying numbers of mutations to display the expected wait time distributions if 50, 100, 200, 300, 400 or 500 mutations occur on internal branches of the phylogeny. Each randomization is run for 10 iterations. The empirical number of S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous mutations observed on internal branches of the phylogeny are indicated.

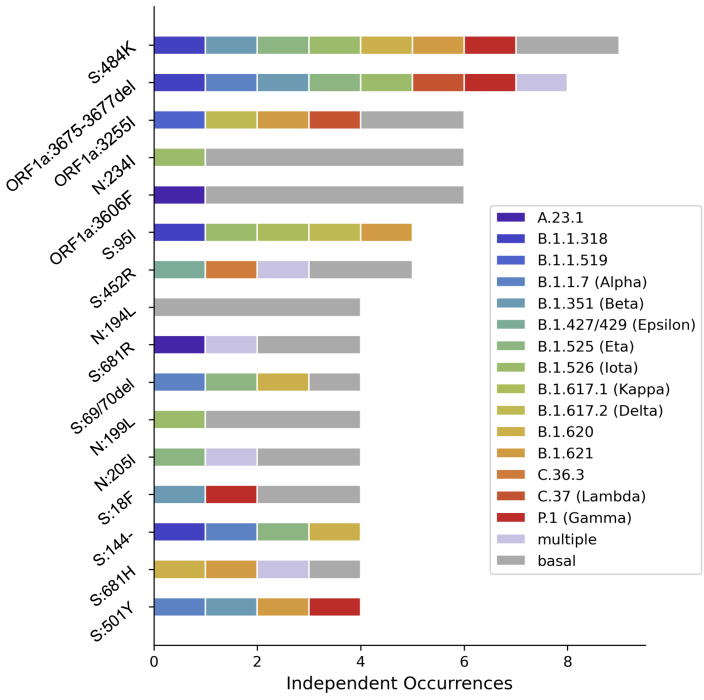

Figure 4.S8: Every occurrence of the 3-amino acid deletion in Nsp6 resulted in an emerging lineage. Every occurrence of the convergently-evolved mutations is colored according to the emerging lineage it occurs at the base of. Multiple emerging lineages descending from the branch a mutation occurs on is represented by light purple.

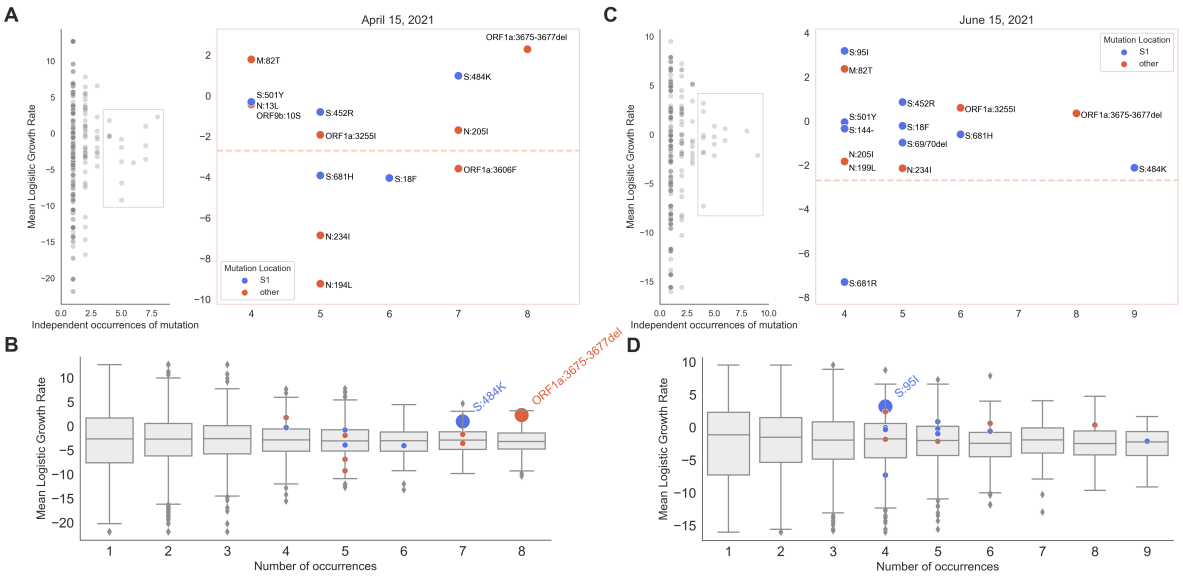

Figure 4.S9: Analyses of convergent evolution shown 1 month before and 1 month after the primary analysis. A) Same as Figure 4.4A, completed using sequences up to April 15, 2021 (1 month before the primary analysis). **B)** Same as Figure 4.4B, completed using sequences up to April 15, 2021. **C)** Same as Figure 4.4A, completed using sequences up to June 15, 2021 (1 month after the primary analysis). **D)** Same as Figure 4.4B, completed using sequences up to June 15, 2021.

Chapter 5

Conclusions

Adaptive evolution leaves distinctive marks on genetic sequences. However, parsing a trail of positive selection from effects of random mutation and neutral evolution can be challenging. In the preceding chapters, I have shown examples of how these marks of adaptive evolution can be identified from the genomes of RNA viruses. In RNA viruses, analyses of adaptive evolution can be confounded by rapid population growth and high mutation rates. Additionally, most established methods are not sensitive enough to pick up signals of directional evolution on the short evolutionary timescales that are relevant for controlling human pathogenic RNA viruses. In this dissertation, each chapter employs different methods to identify adaptive evolution in order to address considerations specific to the evolutionary context of the virus.

Chapter 2 investigates the adaptive evolution of influenza H3N2 during egg-passaging: the main method for growing influenza vaccine virus in the United States. As a consequence of vaccine production, many different human H3N2 isolates have been grown in eggs over the past >50 years of vaccine production, and hundreds of them have been sequenced after egg-passaging. In a sense, this represents an evolution experiment where different human H3N2 strains have been subjected to the same selective pressure (growth in chicken eggs) repeatedly. Thus, much can be learned about how human H3N2 adapts to replication in avian cells through convergent patterns of hundreds of egg-passaged isolates, eliminating the stochasticity of one particular adaptive walk.

In Chapter 2, I built a tree using sequences from egg-passaged and non-egg-passaged viruses, and used phylogenetic inference to identify mutations that repeatedly occurred during egg-passaging. Using enrichment analyses, I showed that certain mutations co-occur more or less frequently than would be expected. The epistatic interactions revealed by this

analysis begin to illuminate the fitness landscape, showing that at least two fitness peaks exist, separated by reciprocal sign epistasis between HA positions 186 and 194, the two predominant sites of egg-adaptation. Through inference of titer values, Chapter 2 shows that adaptive walks up the 194 peak alter antigenicity more than egg-adaptation with a mutation at 186. This suggests that engineering the 186G mutation into candidate vaccine strains might reduce the potential antigenic impact of egg-adaptation during vaccine production. Many of the findings and conclusions of Chapter 2 were also confirmed by structural studies of HA and reported by Wu et al (Wu et al., 2019). This validates the phylogenetic and sequence-based methodology employed in Chapter 2.

While Chapter 2 takes an ad hoc approach to identifying specific adaptive mutations that capitalizes on repetition of the same selective pressure on un-adapted viruses, this approach is not useful for a natural population of viruses circulating in humans. If repeated egg-passaging of H3N2 represents replicates of an evolutionary experiment, the natural evolution of a virus is just a single run of the experiment. Relying only on convergent evolution to identify adaptive evolution during natural history of a virus is simultaneously too insensitive and subject to high rates of false positives. Instead, alternate methods that account for phylogenetic relationships and compare observed evolution to neutral expectations can be employed. In Chapter 3, I use a handful of these methods to address whether or not seasonal coronaviruses undergo adaptive evolution and to quantify this adaptation relative to the better-studied influenza viruses.

There are four seasonal coronaviruses that have been endemic in the human population for 10's to 100's of years. Though these coronaviruses regularly reinfect people, it was not known whether this is due to waning immunity, adaptive evolution of the virus, or both. Chapter 3 investigates the possibility that these viruses evolve adaptively using measures of nonsynonymous and synonymous divergence, a version of the McDonald-Kreitman test tailored to RNA virus evolution, and the phylogenetic shape as given by TMRCA. All of these methods show evidence that OC43 and 229E are evolving adaptively in the S1 domain

of spike, which binds host receptors. Chapter 3 shows that this evolution is occurring at roughly the same rate as influenza B viruses, and the fact that this adaptive evolution is localized to S1 is highly suggestive of antigenic evolution. In 229E, antigenic evolution was confirmed experimentally (Eguia et al., 2021).

Interestingly, Chapter 3 shows a lack of evidence for similar adaptive evolution in NL63. Future work to experimentally confirm that NL63 does not evolve antigenically would offer a very interesting point of comparison between NL63 and the relatively closely-related alphacoronavirus 229E. While it is well-known that some viruses (like influenza) evolve antigenically and others (like measles) do not, 229E and NL63 would be the first known example of two members of the same viral genus that operate differently with regards to antigenic evolution. This is particularly interesting because it could reveal viral characteristics that enable recurrent immune evasion through evolution of the receptor-binding protein subunit.

The human humoral immune response against measles generates antibodies directed against many (∼8) co-dominant epitopes on the measles H protein (Muñoz-Alía et al., 2021). In this virus, lack of antigenic evolution has been attributed to evolutionary inaccessibility of the necessary mutations: mutations within H are not well tolerated (Fulton et al., 2015) and simultaneous disruption of a minimum of 5 of epitopes is required to escape polyclonal sera (Muñoz-Alía et al., 2021). In contrast, influenza HA is very tolerant of mutation (Thyagarajan and Bloom, 2014; Fulton et al., 2015) and can evade polyclonal antibodies with a single amino acid substitution (Koel et al., 2013; Huang et al., 2015). However, measles H and influenza HA are evolutionary distant surface proteins of different viruses with different lifecycles and strategies. A comparison of NL63 and 229E, which share 64.07% identity in spike (Devi and Chaitanya, 2021), could illuminate how specific genetic or environmental differences contribute to the propensity for antigenic evolution in the context of two very similar viruses. Understanding the factors dictating why some viruses evolve antigenically while others do not could enable us to better predict the evolutionary potential of emerging viruses. This, in turn, would allow more effective control of these viruses through proactive

measures to match vaccines to evolutionarily-dominant variants.

For instance, when SARS-CoV-2 entered the human population and became a pandemic in early 2020, little was known about its adaptive potential. At this time, it was commonly touted that coronaviruses do not evolve antigenically, and there was little existing literature about the adaptive potential of coronaviruses (Ren et al., 2015; Chibo and Birch, 2006). However, as evidence of adaptive evolution in seasonal coronavirus spike S1 mounted, so did the circulating variants of SARS-CoV-2 with differing phenotypic properties (Konings et al., 2021) (`https://nextstrain.org/ncov/gisaid/global`). Whether or not these variants represented positively-selected substitutions makes a huge difference in effective suppression of viral transmission. However, SARS-CoV-2 was in its evolutionary infancy, making it especially hard to pinpoint signals of adaptive evolution in a sea of random mutations. The standard methods for this (like those used in Chapter 3) are unable to detect early stages of adaptive evolution when positively-selected residues are still far from fixation.

Chapter 4 presents a new method for identifying adaptive evolution that is able to work on short evolutionary timescales, and draws power from depth of temporal and geographic sampling. This method takes an intuitive approach to this problem by looking for correlations between evolutionarily successful viruses and genomic regions rich in protein-coding changes. Successful clades of viruses are determined by rapid growth rates, which are estimated from a time-resolved phylogeny. Linear regression between a clade's growth rate and the accumulation of different types of mutations is then used to assess the correlation between protein-coding mutations and evolutionary success. Chapter 4 shows that mutations within spike S1, ORF7a, and Nsp6 strongly correlate with clade success. Specific substitutions within these genes as well as Nsp4, N, and M convergently arise many times in the initial year and a half or SARS-CoV-2 evolution and are associated with successful clades each time they do. Many of these changes likely represent adaptations to a new host, including improvements in receptor-binding, replication, and innate immune antagonism.

The majority of adaptive evolution of SARS-CoV-2 is concentrated in the receptor-

binding subunit S1, demonstrating the importance of fine-tuning cell entry after a host jump. However, Chapter 4 suggests that that positively-selected mutations in S1 didn't start amassing until 8-10 months after the start of the pandemic, indicating some change in the fitness landscape. Late 2020 appears to be an inflection point, after which the pace of adaptive evolution in S1 grows steadily to roughly 4 times greater than that of the influenza H3N2 HA1 subunit. It will be interesting to monitor the pace and genomic locations of adaptive evolution as time passes, rates of infection- and vaccine-induced immunity increase, and SARS-CoV-2 becomes endemic. The wealth of sequencing data that continues to amass is an unprecedented resource for studying the dynamics of adaptive evolution of an emerging virus.

As demonstrated in Figure 3.S7, dense temporal sequencing is crucial for identifying adaptive evolution. A concerted effort to sequence other circulating human RNA viruses will enable us to describe and quantify adaptive evolution these viruses using current methods. A more complete understanding of the adaptive capacity of different types of viruses would help to direct surveillance for viruses with pandemic potential and customize prevention and control of existing human pathogens. However, even given millions of genome sequences, it can be hard to disambiguate shared ancestry and stochastic effects from true epistasis. And, without accounting for epistatic effects, there will always be some error when inferring the fitness of a virus from its constituent single mutations or when trying to distill the adaptive effects of individual mutations from an entire genome. Thus, the constraints imposed by epistasis may be one of the largest holes in our models that describe and predict adaptive evolution. Empirical measures of the fitness of every possible combination of mutations is not possible because the number of combinations is vastly large. However, it could be possible to gauge the extent of epistatic constraint and map local regions of the fitness landscape by comparing the effects of single mutations on multiple genetic backgrounds. Though it is a lofty goal, a map of genetic constraints would be extremely useful, when paired with experimental measures of fitness and observed natural evolution, for improving evolutionary

predictions.

References

I Barberis, P Myles, S K Ault, N L Bragazzi, and M Martini. History and evolution of influenza control through vaccination: from the first monovalent vaccine to universal vaccines. *J. Prev. Med. Hyg.*, 57(3):E115–E120, September 2016.

Subrata Barman, John Franks, Jasmine C Turner, Sun-Woo Yoon, Robert G Webster, and Richard J Webby. Egg-adaptive mutations in H3N2v vaccine virus enhance egg-based production without loss of antigenicity or immunogenicity. *Vaccine*, 33(28):3186–3192, June 2015.

Ian G Barr, Ruben O Donis, Jacqueline M Katz, John W McCauley, Takato Odagiri, Heidi Trusheim, Theodore F Tsai, and David E Wentworth. Cell culture-derived influenza vaccines in the severe 2017–2018 epidemic season: a step towards improved influenza vaccine effectiveness. *npj Vaccines*, 3(1):44, October 2018.

Trevor Bedford, Sarah Cobey, and Mercedes Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol. Biol.*, 11:220, July 2011.

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, February 2014.

Edward A Belongia, Melissa D Simpson, Jennifer P King, Maria E Sundaram, Nicholas S Kelley, Michael T Osterholm, and Huong Q McLean. Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies. *Lancet Infect. Dis.*, 16(8):942–951, August 2016.

Domenico Benvenuto, Silvia Angeletti, Marta Giovanetti, Martina Bianchi, Stefano Pascarella, Roberto Cauda, Massimo Ciccozzi, and Antonio Cassone. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural protein 6 (NSP6) could affect viral autophagy. *J. Infect.*, 81(1):e24–e27, July 2020.

Samir Bhatt, Aris Katzourakis, and Oliver G Pybus. Detecting natural selection in RNA virus populations using sequence summary statistics. *Infect. Genet. Evol.*, 10(3):421–430, April 2010.

Samir Bhatt, Edward C Holmes, and Oliver G Pybus. The genomic rate of molecular adaptation of the human influenza a virus. *Mol. Biol. Evol.*, 28(9):2443–2451, September 2011.

Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K Mendes, Nicola F Müller, Huw A Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J Drummond.

BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.*, 15(4):e1006650, April 2019.

Rena Brauer and Peter Chen. Influenza virus propagation in embryonated chicken eggs. *J. Vis. Exp.*, (97), March 2015.

Finlay Campbell, Brett Archer, Henry Laurenson-Schafer, Yuka Jinnai, Franck Konings, Neale Batra, Boris Pavlin, Katelijn Vandemaele, Maria D Van Kerkhove, Thibaut Jombart, Oliver Morgan, and Olivier le Polain de Waroux. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at june 2021. *Eurosurveillance*, 26 (24):2100509, June 2021.

Centers for Disease Control and Prevention. Flu vaccine and people with egg allergies. `https://www.cdc.gov/flu/prevent/egg-allergies.htm`, September 2021. Accessed: 2021-11-20.

Zhongying Chen, Helen Zhou, and Hong Jin. The impact of key amino acid substitutions in the hemagglutinin of influenza a (H3N2) viruses on vaccine production and antibody response. *Vaccine*, 28(24):4079–4085, May 2010.

Doris Chibo and Chris Birch. Analysis of human coronavirus 229E spike and nucleoprotein genes demonstrates genetic drift between chronologically distinct strains. *J. Gen. Virol.*, 87(Pt 5):1203–1208, May 2006.

Sarah Cobey, Sigrid Gouma, Kaela Parkhouse, Benjamin S Chambers, Hildegund C Ertl, Kenneth E Schmader, Rebecca A Halpin, Xudong Lin, Timothy B Stockwell, Suman R Das, Emily Landon, Vera Tesic, Ilan Youngster, Benjamin A Pinsky, David E Wentworth, Scott E Hensley, and Yonatan H Grad. Poor immunogenicity, not vaccine strain egg adaptation, may explain the low H3N2 influenza vaccine effectiveness in 2012–2013. *Clin. Infect. Dis.*, 67(3):327–333, July 2018.

Arpita Devi and Nyshadham S N Chaitanya. In silico designing of multi-epitope vaccine construct against human coronavirus infections. *J. Biomol. Struct. Dyn.*, 39(18):6903–6917, November 2021.

J W Drake. Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. U. S. A.*, 90(9):4171–4175, May 1993.

Arthur W D Edridge, Joanna Kaczorowska, Alexis C R Hoste, Margreet Bakker, Michelle Klein, Katherine Loens, Maarten F Jebbink, Amy Matser, Cormac M Kinsella, Paloma Rueda, Margareta Ieven, Herman Goossens, Maria Prins, Patricia Sastre, Martin Deijs, and Lia van der Hoek. Seasonal coronavirus protective immunity is short-lasting. *Nat. Med.*, September 2020.

Rachel Eguia, Katharine H D Crawford, Terry Stevens-Ayers, Laurel Kelnhofer-Millevolte, Alexander L Greninger, Janet A Englund, Michael J Boeckh, and Jesse D Bloom. A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog.*, 17: e1009453, 2021.

European Centre for Disease Prevention and Control. Influenza virus characterisation, summary europe. Technical report, European Centre for Disease Prevention and Control, Stockholm, February 2015.

Benjamin O Fulton, David Sachs, Shannon M Beaty, Sohui T Won, Benhur Lee, Peter Palese, and Nicholas S Heaton. Mutational analysis of measles virus suggests constraints on antigenic variation of the glycoproteins. *Cell Rep.*, 11(9):1331–1338, June 2015.

Allison J Greaney, Andrea N Loes, Katharine H D Crawford, Tyler N Starr, Keara D Malone, Helen Y Chu, and Jesse D Bloom. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe*, 29(3):463–476.e6, March 2021.

James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018.

D Hamre and M Beem. Virologic studies of acute respiratory disease in young adults. v. coronavirus 229E infections during six years of surveillance. *Am. J. Epidemiol.*, 96(2): 94–106, August 1972.

Alfred T Harding and Nicholas S Heaton. Efforts to improve the seasonal influenza vaccine. *Vaccines (Basel)*, 6(2), March 2018.

Terho Heikkinen and Asko Järvinen. The common cold. *Lancet*, 361(9351):51–59, January 2003.

Scott E Hensley, Suman R Das, Adam L Bailey, Loren M Schmidt, Heather D Hickman, Akila Jayaraman, Karthik Viswanathan, Rahul Raman, Ram Sasisekharan, Jack R Bennink, and Jonathan W Yewdell. Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. *Science*, 326(5953):734–736, October 2009.

Markus Hoffmann, Hannah Kleine-Weber, and Stefan Pöhlmann. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell*, 78(4):779–784.e5, May 2020.

Heike Hofmann, Graham Simmons, Andrew J Rennekamp, Chawaree Chaipan, Thomas Gramberg, Elke Heck, Martina Geier, Anja Wegele, Andrea Marzi, Paul Bates, and Stefan Pöhlmann. Highly conserved regions within the spike proteins of human coronaviruses 229E and NL63 determine recognition of their respective cellular receptors. *J. Virol.*, 80 (17):8639–8652, September 2006.

J J Holland, J C De La Torre, and D A Steinhauer. RNA virus populations as quasispecies. *Curr. Top. Microbiol. Immunol.*, 176:1–20, 1992.

Chung-Chau Hon, Tsan-Yuk Lam, Zheng-Li Shi, Alexei J Drummond, Chi-Wai Yip, Fanya Zeng, Pui-Yi Lam, and Frederick Chi-Ching Leung. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J. Virol.*, 82(4):1819–1826, February 2008.

Kuan-Ying A Huang, Pramila Rijal, Lisa Schimanski, Timothy J Powell, Tzou-Yien Lin, John W McCauley, Rodney S Daniels, and Alain R Townsend. Focused antibody response to influenza linked to antigenic drift. *J. Clin. Invest.*, 125(7):2631–2645, July 2015.

John Huddleston, James Hadfield, Thomas R Sibley, Jover Lee, Kairsten Fay, Misja Ilcisin, Elias Harkins, Trevor Bedford, Richard A Neher, and Emma B Hodcroft. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *J Open Source Softw*, 6(57), January 2021.

Ruben J G Hulswit, Yifei Lang, Mark J G Bakkers, Wentao Li, Zeshi Li, Arie Schouten, Bram Ophorst, Frank J M van Kuppeveld, Geert-Jan Boons, Berend-Jan Bosch, Eric G Huizinga, and Raoul J de Groot. Human coronaviruses OC43 and HKU1 bind to 9-o-acetylated sialic acids via a conserved receptor-binding site in spike protein domain a. *Proc. Natl. Acad. Sci. U. S. A.*, 116(7):2681–2690, February 2019.

Aeron C Hurt. The epidemiology and spread of drug resistant human influenza viruses. *Curr. Opin. Virol.*, 8:22–29, October 2014.

Abbas Jariani, Christopher Warth, Koen Deforche, Pieter Libin, Alexei J Drummond, Andrew Rambaut, Frederick A Matsen, Iv, and Kristof Theys. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evol*, 5(1):vez003, January 2019.

Irwin Jungreis, Rachel Sealfon, and Manolis Kellis. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 sarbecovirus genomes. *Nat. Commun.*, 12(1):1–20, May 2021.

Sisi Kang, Mei Yang, Suhua He, Yueming Wang, Xiaoxue Chen, Yao-Qing Chen, Zhongsi Hong, Jing Liu, Guanmin Jiang, Qiuyue Chen, Ziliang Zhou, Zhechong Zhou, Zhaoxia Huang, Xi Huang, Huanhuan He, Weihong Zheng, Hua-Xin Liao, Fei Xiao, Hong Shan, and Shoudeng Chen. A SARS-CoV-2 antibody curbs viral nucleocapsid protein-induced complement hyperactivation. *Nat. Commun.*, 12(1):1–11, May 2021.

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 Res.*, 30(14):3059–3066, July 2002.

Kathryn E Kistler and Trevor Bedford. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. *Elife*, 10, January 2021.

Patience K Kiyuka, Charles N Agoti, Patrick K Munywoki, Regina Njeru, Anne Bett, James R Otieno, Grieven P Otieno, Everlyn Kamau, Taane G Clark, Lia van der Hoek, Paul Kellam, D James Nokes, and Matthew Cotten. Human coronavirus NL63 molecular epidemiology and evolutionary patterns in rural coastal kenya. *J. Infect. Dis.*, 217(11): 1728–1739, May 2018.

Björn 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):976–979, November 2013.

Kenichi Komabayashi, Yohei Matoba, Shizuka Tanaka, Junji Seto, Yoko Aoki, Tatsuya Ikeda, Yoshitaka Shimotai, Yoko Matsuzaki, Tsutomu Itagaki, and Katsumi Mizuta. Longitudinal epidemiology of human coronavirus OC43 in yamagata, japan, 2010–2017: Two groups based on spike gene appear one after another. *J. Med. Virol.*, 7:825, August 2020.

Frank Konings, Mark D Perkins, Jens H Kuhn, Mark J Pallen, Erik J Alm, Brett N Archer, Amal Barakat, Trevor Bedford, Jinal N Bhiman, Leon Caly, Lisa L Carter, Anne Cullinane, Tulio de Oliveira, Julian Druce, Ihab El Masry, Roger Evans, George F Gao, Alexander E Gorbalenya, Esther Hamblion, Belinda L Herring, Emma Hodcroft, Edward C Holmes, Manish Kakkar, Shagun Khare, Marion P G Koopmans, Bette Korber, Juliana Leite, Duncan MacCannell, Marco Marklewitz, Sebastian Maurer-Stroh, Jairo Andres Mendez Rico, Vincent J Munster, Richard Neher, Bas Oude Munnink, Boris I Pavlin, Malik Peiris, Leo Poon, Oliver Pybus, Andrew Rambaut, Paola Resende, Lorenzo Subissi, Volker Thiel, Suxiang Tong, Sylvie van der Werf, Anne von Gottberg, John Ziebuhr, and Maria D Van Kerkhove. SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nature Microbiology*, 6(7):821–823, June 2021.

Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.*, 23(10):1891–1901, October 2006.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

Florian Krammer. SARS-CoV-2 vaccines in development, 2020.

Sergey Kryazhimskiy and Joshua B Plotkin. The population genetics of dN/dS. *PLoS Genet.*, 4(12), 2008.

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, 2015.

Susanna K P Lau, Paul Lee, Alan K L Tsang, Cyril C Y Yip, Herman Tse, Rodney A Lee, Lok-Yee So, Yu-Lung Lau, Kwok-Hung Chan, Patrick C Y Woo, and Kwok-Yung Yuen. Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. *J. Virol.*, 85(21):11325–11337, November 2011.

Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C Hurt, Seema S Lakdawala, Scott E Hensley, and Jesse D Bloom. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife*, 8, August 2019.

Min Z Levine, Emily T Martin, Joshua G Petrie, Adam S Lauring, Crystal Holiday, Stacie Jefferson, William J Fitzsimmons, Emileigh Johnson, Jill M Ferdinands, and Arnold S Monto. Antibodies against egg- and cell-grown influenza A(H3N2) viruses in adults hospitalized during the 2017-2018 season. October 2018.

Fang Li. Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol*, 3(1):237–261, September 2016.

Qianqian Li, Jiajing Wu, Jianhui Nie, Li Zhang, Huan Hao, Shuo Liu, Chenyan Zhao, Qi Zhang, Huan Liu, Lingling Nie, Haiyang Qin, Meng Wang, Qiong Lu, Xiaoyu Li, Qiyu Sun, Junkai Liu, Linqi Zhang, Xuguang Li, Weijin Huang, and Youchun Wang. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*, 182 (5):1284–1294.e9, September 2020.

Zhijie Li, Aidan Ca Tomlinson, Alan Hm Wong, Dongxia Zhou, Marc Desforges, Pierre J Talbot, Samir Benlekbir, John L Rubinstein, and James M Rini. The human coronavirus HCoV-229E s-protein structure and receptor binding. *Elife*, 8, October 2019.

Yi Pu Lin, Victoria Gregory, Patrick Collins, Johannes Kloess, Stephen Wharton, Nicholas Cattle, Angie Lackenby, Rodney Daniels, and Alan Hay. Neuraminidase receptor binding variants of human influenza A(H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: a role in virus attachment? *J. Virol.*, 84(13):6769–6781, July 2010.

Ding X Liu, Jia Q Liang, and To S Fung. Human Coronavirus-229E, -OC43, -NL63, and -HKU1, 2020a.

Lihong Liu, Pengfei Wang, Manoj S Nair, Jian Yu, Micah Rapp, Qian Wang, Yang Luo, Jasper F-W Chan, Vincent Sahi, Amir Figueroa, Xinzheng V Guo, Gabriele Cerutti, Jude Bimela, Jason Gorman, Tongqing Zhou, Zhiwei Chen, Kwok-Yung Yuen, Peter D Kwong, Joseph G Sodroski, Michael T Yin, Zizhang Sheng, Yaoxing Huang, Lawrence Shapiro, and David D Ho. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature*, 584(7821):450–456, August 2020b.

Yang Liu, Jianying Liu, Bryan A Johnson, Hongjie Xia, Zhiqiang Ku, Craig Schindewolf, Steven G Widen, Zhiqiang An, Scott C Weaver, Vineet D Menachery, Xuping Xie, and Pei-Yong Shi. Delta spike P681R mutation enhances SARS-CoV-2 fitness over alpha variant. *bioRxiv*, September 2021.

Ben Longdon, Michael A Brockhurst, Colin A Russell, John J Welch, and Francis M Jiggins. The evolution and genetics of virus host shifts. *PLoS Pathog.*, 10(11):e1004395, November 2014.

Bin Lu, Helen Zhou, Dan Ye, George Kemble, and Hong Jin. Improvement of influenza A/Fujian/411/02 (H3N2) virus growth in embryonated chicken eggs by balancing the hemagglutinin and neuraminidase activities, using reverse genetics. *J. Virol.*, 79(11): 6763–6771, June 2005.

Bailey Lubinski, Laura E Frazier, My T Phan, V, Daniel L Bugembe, Tiffany Tang, Susan Daniel, Matthew Cotten, Javier A Jaimes, and Gary R Whittaker. Spike protein cleavage-activation mediated by the SARS-CoV-2 P681R mutation: a case-study from its first appearance in variant of interest (VOI) a.23.1 identified in uganda. *bioRxiv*, July 2021.

Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507 (7490):57–61, March 2014.

Darren P Martin, Steven Weaver, Houriiyah Tegally, Emmanuel James San, Stephen D Shank, Eduan Wilkinson, Alexander G Lucaci, Jennifer Giandhari, Sureshnee Naidoo, Yeshnee Pillay, Lavanya Singh, Richard J Lessells, Ravindra K Gupta, Joel O Wertheim, Anton Nekturenko, Ben Murrell, Gordon W Harkins, Philippe Lemey, Oscar A MacLean, David L Robertson, Tulio de Oliveira, and Sergei L Kosakovsky Pond. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*, September 2021.

Kevin R McCarthy, Linda J Rennick, Sham Nambulli, Lindsey R Robinson-McCarthy, William G Bain, Ghady Haidar, and W Paul Duprex. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science*, 371(6534):1139–1142, March 2021.

J H McDonald and M Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652–654, June 1991.

Kenneth McIntosh. Coronaviruses: A comparative review. In *Current Topics in Microbiology and Immunology / Ergebnisse der Mikrobiologie und Immunitätsforschung*, pages 85–129. Springer Berlin Heidelberg, 1974.

Arnold S Monto and Sook K Lim. The tecumseh study of respiratory illness. VI. frequency of and relationship between outbreaks of coronavims infection. *J. Infect. Dis.*, 129(3): 271–276, March 1974.

Nicola F Müller, Kathryn E Kistler, and Trevor Bedford. Recombination patterns in coronaviruses. April 2021.

Miguel Ángel Muñoz-Alía, Rebecca A Nace, Lianwen Zhang, and Stephen J Russell. Serotypic evolution of measles virus is constrained by multiple co-dominant B cell epitopes on its surface glycoproteins. *Cell Rep Med*, 2(4):100225, April 2021.

Ben Murrell, Joel O Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, and Sergei L Kosakovsky Pond. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.*, 8(7):e1002764, July 2012.

Sabine Nakowitsch, Andrea M Waltenberger, Nina Wressnigg, Nicole Ferstl, Andrea Triendl, Bettina Kiefmann, Emanuele Montomoli, Giulia Lapini, Maria Sergeeva, Thomas Muster, and Julia R Romanova. Egg- or cell culture-derived hemagglutinin mutations impair virus stability and antigen content of inactivated influenza vaccines, 2014.

Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. pages 1701–1709, 2016.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274, January 2015.

Eri Nobusawa and Katsuhiko Sato. Comparison of the mutation rates of human influenza a and B viruses. *J. Virol.*, 80(7):3675–3678, April 2006.

Lauren Parker, Stephen A Wharton, Stephen R Martin, Karen Cross, Yipu Lin, Yan Liu, Ten Feizi, Rodney S Daniels, and John W McCauley. Effects of egg-adaptation on receptor-binding and antigenic properties of recent influenza a (H3N2) vaccine viruses. *J. Gen. Virol.*, 97(6):1333–1344, June 2016.

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

Alexander O Pasternak, Willy J M Spaan, and Eric J Snijder. Nidovirus transcription: how to make sense...? *J. Gen. Virol.*, 87(6):1403–1421, 2006.

Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–619, May 2008.

Sylvia E Reed. The behaviour of recent isolates of human respiratory coronavirus in vitro and in volunteers: Evidence of heterogeneity among 229e-related strains. *J. Med. Virol.*, 13(2):179–192, 1984.

Lili Ren, Yue Zhang, Jianguo Li, Yan Xiao, Jing Zhang, Ying Wang, Lan Chen, Gláucia Paranhos-Baccalà, and Jianwei Wang. Genetic drift of human coronavirus OC43 spike gene during adaptive evolution. *Sci. Rep.*, 5:11451, June 2015.

Nash D Rochman, Yuri I Wolf, Guilhem Faure, Pascal Mutz, Feng Zhang, and Eugene V Koonin. Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.*, 118(29), July 2021.

Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*, 4(1):vex042, January 2018.

Rafael Sanjuán and Antonio V Bordería. Interplay between RNA structure and protein evolution in HIV-1. *Mol. Biol. Evol.*, 28(4):1333–1338, April 2011.

Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22(13), March 2017.

K L Simonsen, G A Churchill, and C F Aquadro. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141(1):413–429, September 1995.

Danuta M Skowronski and Gaston De Serres. Role of egg-adaptation mutations in low influenza A(H3N2) vaccine effectiveness during the 2012–2013 season, 2018.

Danuta M Skowronski, Naveed Z Janjua, Gaston De Serres, Suzana Sabaiduc, Alireza Eshaghi, James A Dickinson, Kevin Fonseca, Anne-Luise Winter, Jonathan B Gubbay, Mel Krajden, Martin Petric, Hugues Charest, Nathalie Bastien, Trijntje L Kwindt, Salaheddin M Mahmud, Paul Van Caeseele, and Yan Li. Low 2012–13 influenza vaccine effectiveness associated with mutation in the Egg-Adapted H3N2 vaccine strain not antigenic drift in circulating viruses. *PLoS One*, 9(3):e92153, March 2014.

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.

Nick G C Smith and Adam Eyre-Walker. Adaptive protein evolution in drosophila. *Nature*, 415(6875):1022–1024, February 2002.

Redmond P Smyth, Matteo Negroni, Andrew M Lever, Johnson Mak, and Julia C Kenyon. RNA Structure—A neglected puppet master for the evolution of virus and host immunity. *Front. Immunol.*, 9:2097, 2018.

Eric J Snijder, Ronald W A L Limpens, Adriaan H de Wilde, Anja W M de Jong, Jessika C Zevenhoven-Dobbe, Helena J Maier, Frank F G A Faas, Abraham J Koster, and Montserrat Bárcena. A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. *PLoS Biol.*, 18(6):e3000715, June 2020.

Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine H D Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, Neil P King, David Veesler, and Jesse D Bloom. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.e20, September 2020.

Yvonne C F Su, Justin Bahl, Udayan Joseph, Ka Man Butt, Heidi A Peck, Evelyn S C Koay, Lynette L E Oon, Ian G Barr, Dhanasekaran Vijaykrishna, and Gavin J D Smith. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat. Commun.*, 6(1):1–13, August 2015.

F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, November 1989.

Kristof Theys, Pieter Libin, Andrea-Clemencia Pineda-Peña, Ann Nowé, Anne-Mieke Vandamme, and Ana B Abecasis. The impact of HIV-1 within-host evolution on transmission dynamics. *Curr. Opin. Virol.*, 28:92–101, February 2018.

Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3, July 2014.

Yatish Turkahia, Bryan Thornlow, Angie Hinrichs, Jakob McBroome, Nicolas Ayala, Cheng Ye, Nicola De Maio, David Haussler, Robert Lanfear, and Russell Corbett-Detig. Pandemic-Scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. August 2021.

Lucy van Dorp, Mislav Acman, Damien Richard, Liam P Shaw, Charlotte E Ford, Louise Ormond, Christopher J Owen, Juanita Pang, Cedric C S Tan, Florencia A T Boshier, Arturo Torres Ortiz, and François Balloux. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.*, 83:104351, September 2020.

Leen Vijgen, Philippe Lemey, Els Keyaerts, and Marc Van Ranst. Genetic variability of human respiratory coronavirus OC43. *J. Virol.*, 79(5):3223–4; author reply 3224–5, March 2005.

Erik M Volz, Katia Koelle, and Trevor Bedford. Viral phylodynamics. *PLoS Comput. Biol.*, 9(3):e1002947, March 2013.

Zeng Wang, Huanliang Yang, Yan Chen, Shiyu Tao, Liling Liu, Huihui Kong, Shujie Ma, Fei Meng, Yasuo Suzuki, Chuanling Qiao, and Hualan Chen. A Single-Amino-Acid substitution at position 225 in hemagglutinin alters the transmissibility of eurasian Avian-Like H1N1 swine influenza virus in guinea pigs. *J. Virol.*, 91(21), November 2017.

Zijun Wang, Fabian Schmidt, Yiska Weisblum, Frauke Muecksch, Christopher O Barnes, Shlomo Finkin, Dennis Schaefer-Babajew, Melissa Cipolla, Christian Gaebler, Jenna A Lieberman, Thiago Y Oliveira, Zhi Yang, Morgan E Abernathy, Kathryn E Huey-Tubman, Arlene Hurley, Martina Turroja, Kamille A West, Kristie Gordon, Katrina G Millard, Victor Ramos, Justin Da Silva, Jianliang Xu, Robert A Colbert, Roshni Patel, Juan Dizon, Cecille Unson-O'Brien, Irina Shimeliovich, Anna Gazumyan, Marina Caskey, Pamela J Bjorkman, Rafael Casellas, Theodora Hatziioannou, Paul D Bieniasz, and Michel C Nussenzweig. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature*, 592(7855):616–622, April 2021.

T Watabe and H Kishino. Structural considerations in the fitness landscape of a virus, 2010.

Steven Weaver, Stephen D Shank, Stephanie J Spielman, Michael Li, Spencer V Muse, and Sergei L Kosakovsky Pond. Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes, 2018.

Linda Widjaja, Natalia Ilyushina, Robert G Webster, and Richard J Webby. Molecular changes associated with adaptation of human influenza a virus in embryonated chicken eggs. *Virology*, 350(1):137–145, June 2006.

Scott Williamson. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.*, 20(8):1318–1325, August 2003.

Jeroen Witteveldt, Richard Blundell, Joris J Maarleveld, Nora McFadden, David J Evans, and Peter Simmonds. The influence of viral RNA secondary structure on interactions with innate host cell defences. *Nucleic Acids Res.*, 42(5):3314–3329, March 2014.

Yuri I Wolf, Cecile Viboud, Edward C Holmes, Eugene V Koonin, and David J Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza a virus. *Biol. Direct*, 1:34, October 2006.

Patrick C Y Woo, Susanna K P Lau, Yi Huang, and Kwok-Yung Yuen. Coronavirus diversity, phylogeny and interspecies jumping. *Exp. Biol. Med.*, 234(10):1117–1127, October 2009.

World Health Organization. Seasonal influenza fact sheet. `https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)`. Accessed: 2021-11-17.

Worldwide Influenza Center: Francis Crick Institute. WHO vaccine recommendation meeting for the southern hemisphere 2019 influenza season. Technical report, September 2018.

Sewall Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. 1932.

Nicholas C Wu, Seth J Zost, Andrew J Thompson, David Oyen, Corwin M Nycholat, Ryan McBride, James C Paulson, Scott E Hensley, and Ian A Wilson. A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathog.*, 13(10): 1–17, 2017.

Nicholas C Wu, Huibin Lv, Andrew J Thompson, Douglas C Wu, Wilson W S Ng, Rameshwar U Kadam, Chih-Wei Lin, Corwin M Nycholat, Ryan McBride, Weiwen Liang, James C Paulson, Chris K P Mok, and Ian A Wilson. Preventing an antigenically disruptive mutation in Egg-Based H3N2 seasonal influenza vaccines by mutational incompatibility. *Cell Host Microbe*, May 2019.

Hongjie Xia, Zengguo Cao, Xuping Xie, Xianwen Zhang, John Yun-Chung Chen, Hualei Wang, Vineet D Menachery, Ricardo Rajsbaum, and Pei-Yong Shi. Evasion of type I interferon by SARS-CoV-2. *Cell Rep.*, 33(1):108234, October 2020.

Jia Xue, Benjamin S Chambers, Scott E Hensley, and Carolina B López. Propagation and characterization of influenza virus stocks that lack high levels of defective viral genomes and hemagglutinin mutations. *Front. Microbiol.*, 7:326, March 2016.

Seiya Yamayoshi and Yoshihiro Kawaoka. Current and future influenza vaccines. *Nat. Med.*, 25(2):212–220, February 2019.

Lei Yang, Yanhui Cheng, Xiang Zhao, Hejiang Wei, Minju Tan, Xiyan Li, Wenfei Zhu, Weijuan Huang, Wenbing Chen, Jia Liu, Zi Li, Yuelong Shu, and Dayan Wang. Mutations associated with egg adaptation of influenza A(H1N1)pdm09 virus in laboratory based surveillance in china, 2009–2016. *Biosafety and Health*, 1(1):41–45, June 2019.

Ziheng Yang and Rasmus Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, 2000.

Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. Population genomics of intrapatient HIV-1 evolution. *Elife*, 4, December 2015.

Senyan Zhang, Panpan Zhou, Pengfei Wang, Yangyang Li, Liwei Jiang, Wenxu Jia, Han Wang, Angela Fan, Dongli Wang, Xuanling Shi, Xianyang Fang, Michal Hammel, Shuying Wang, Xinquan Wang, and Linqi Zhang. Structural definition of a unique neutralization epitope on the Receptor-Binding domain of MERS-CoV spike glycoprotein, 2018.

Yue Zhang, Jianguo Li, Yan Xiao, Jing Zhang, Ying Wang, Lan Chen, Gláucia Paranhos-Baccalà, Lili Ren, and Jianwei Wang. Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination, 2015.

Haixia Zhou, Yingzhu Chen, Shuyuan Zhang, Peihua Niu, Kun Qin, Wenxu Jia, Baoying Huang, Senyan Zhang, Jun Lan, Linqi Zhang, Wenjie Tan, and Xinquan Wang. Structural definition of a neutralization epitope on the n-terminal domain of MERS-CoV spike glycoprotein, 2019.

Yun Zhu, Changchong Li, Li Chen, Baoping Xu, Yunlian Zhou, Ling Cao, Yunxiao Shang, Zhou Fu, Aihuan Chen, Li Deng, Yixiao Bao, Yun Sun, Limin Ning, Chunyan Liu, Ju Yin, Zhengde Xie, and Kunling Shen. A novel human coronavirus OC43 genotype detected in mainland china. *Emerg. Microbes Infect.*, 7(1):173, October 2018.

Seth J Zost, Kaela Parkhouse, Megan E Gumina, Kangchon Kim, Sebastian Diaz Perez, Patrick C Wilson, John J Treanor, Andrea J Sant, Sarah Cobey, and Scott E Hensley. Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences*, 114(47):201712377, 2017.