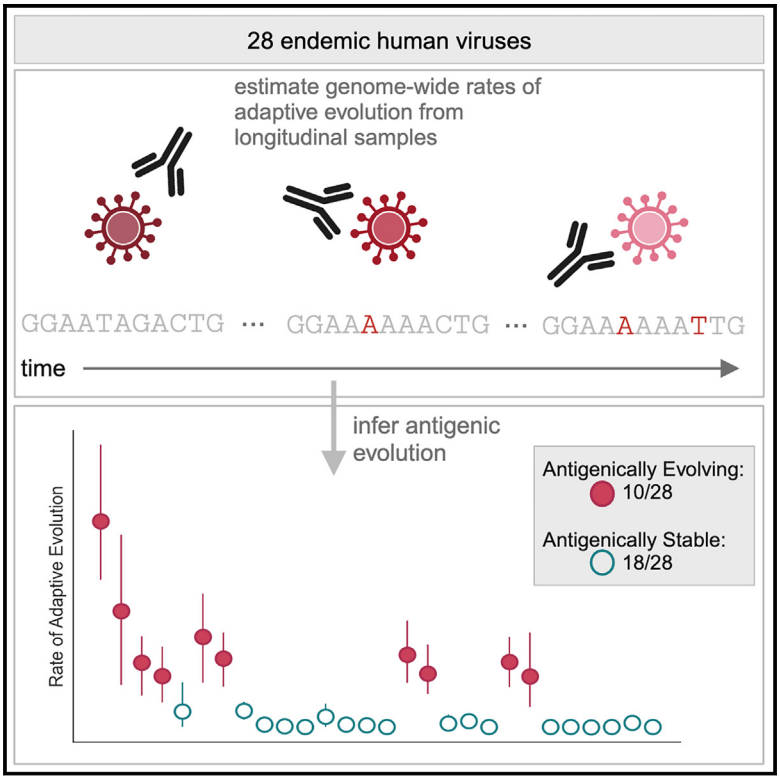
Article

Cell Host & Microbe

An atlas of continuous adaptive evolution in endemic human viruses

Graphical abstract



Authors

Kathryn E. Kistler, Trevor Bedford

Correspondence

kkistler@fredhutch.org

In brief

Kistler and Bedford examine the genomes of 28 human endemic viruses and estimate that 10 of these viruses are undergoing antigenic evolution. This demonstrates that evasion of antibody detection is not an uncommon evolutionary strategy among the viruses that commonly infect humans.

Highlights

- Ongoing adaptive evolution in human endemic viruses is largely in surface proteins
- Immune evasion drives continuous adaptive evolution in many endemic human viruses
- Antigenic evolution occurs in several viral families
- SARS-CoV-2 is accumulating protein-coding changes faster than other endemic viruses

Kistler & Bedford, 2023, Cell Host & Microbe *31*, 1–12
November 8, 2023
https://doi.org/10.1016/j.chom.2023.09.012

CellPress

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

Cell Host & Microbe

CellPress
OPEN ACCESS

Article

An atlas of continuous adaptive evolution in endemic human viruses

Kathryn E. Kistler[1,2,3,*] and Trevor Bedford[1,2]
[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
[2]Howard Hughes Medical Institute, Seattle, WA, USA
[3]Lead contact
*Correspondence: kkistler@fredhutch.org
https://doi.org/10.1016/j.chom.2023.09.012

SUMMARY

Through antigenic evolution, viruses such as seasonal influenza evade recognition by neutralizing antibodies. This means that a person with antibodies well tuned to an initial infection will not be protected against the same virus years later and that vaccine-mediated protection will decay. To expand our understanding of which endemic human viruses evolve in this fashion, we assess adaptive evolution across the genome of 28 endemic viruses spanning a wide range of viral families and transmission modes. Surface proteins consistently show the highest rates of adaptation, and ten viruses in this panel are estimated to undergo antigenic evolution to selectively fix mutations that enable the escape of prior immunity. Thus, antibody evasion is not an uncommon evolutionary strategy among human viruses, and monitoring this evolution will inform future vaccine efforts. Additionally, by comparing overall amino acid substitution rates, we show that SARS-CoV-2 is accumulating protein-coding changes at substantially faster rates than endemic viruses.

INTRODUCTION

Because of their fast mutation rates and high offspring production, many viruses are capable of rapidly evolving to persist and thrive in a changing environment. In the context of human health and disease, this rapid evolution means that viruses from a different host species that cause sporadic human infections can sometimes optimize their cell entry, replication, and immune evasion quickly enough to spread from human to human and become a novel pathogen. Thus, the early stages of a pandemic are often marked by high rates of adaptive evolution, as was noted for the 2009 influenza H1N1 pandemic[1–3] and during the emergence and spread of *variant* viruses during the SARS-CoV-2 pandemic in late 2020 and early 2021.[4]

After this initial adaptation to a new host, some viruses find a niche as endemic viruses, where they are able to infect, replicate in, and transmit between humans without continuous adaptive evolution. However, other endemic viruses continue to evolve adaptively.[5–7] A well-recognized form of this continuing adaptation is antigenic evolution, where the virus and the human adaptive immune system engage in a back-and-forth evolutionary battle—the immune system to neutralize the virus and the virus to evade neutralization. Viruses that evolve antigenically are particularly capable of causing repeat infections and escaping vaccine-mediated immunity.[8] Therefore, understanding which viruses evolve in this manner is highly relevant for managing viral transmission and mitigating human disease.

Antigenic evolution is a well-noted phenomenon of influenza A/H3N2, where this type of evolution necessitates nearly yearly reformulations of the seasonal influenza vaccine.[5,9,10] Serological testing has also demonstrated antigenic evolution in influenza viruses A/H1N1pdm, B/Victoria, and B/Yamagata,[11,12] as well as in seasonal coronavirus 229E[7] and SARS-CoV-2.[13–15] On the contrary, measles[16,17] and influenza C viruses[18] are known to be antigenically stable and do not undergo this mode of continual adaptive evolution. Whether other endemic human pathogenic viruses evolve antigenically is less well understood.

Here, we aim to survey the potential for antigenic evolution or other continuous adaptive evolution across a broad diversity of endemic human viruses. To do this, we use sequencing data to estimate rates of adaptive evolution across each gene in the genome of 28 viruses, which span 10 viral families and a variety of modes of human-to-human transmission. We identify potential antigenically evolving viruses as those with high rates of adaptation in the protein that mediates receptor binding, as this is a primary location of antibody neutralization and the locus of antigenic escape mutations in seasonal influenza viruses,[9,19–21] coronavirus 229E,[7] and SARS-CoV-2.[22–25]

We estimate rates of adaptive evolution from the genetic sequences of viral isolates that have been sampled over time using a McDonald-Kreitman-based method[26,27] that was formulated for analyzing RNA viruses by Williamson[28] and, later, Bhatt et al.[29,30] and then further improved in this manuscript to account for repeated mutations at the same nucleotide position. By estimating rates of adaptation in units of adaptive mutations per codon per year, this method allows us to directly compare adaptive evolution both across the genes of a genome and between different viruses. We find that in addition to seasonal influenza

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

A and B viruses, norovirus, respiratory syncytial virus (RSV-A and -B), two seasonal coronaviruses (229E and OC43-A), and enterovirus D68 all have elevated rates of adaptation in their receptor-binding proteins, indicating potential antigenic evolution in these viruses. Our results not only increase our understanding of ongoing adaptive evolution in current endemic viruses but also provide an expectation of antigenic evolution in other related viruses, including future pandemic viruses. In addition to this manuscript, we have made our results viewable at blab.github.io/atlas-of-viral-adaptation/, where interactive plots allow the user to investigate the results as either a comparison of different viruses or as a comparison across the genome of a single virus.

RESULTS

An extension of the McDonald-Kreitman method for estimating rates of adaptation in viral genomes

Viruses that undergo antigenic evolution repeatedly evade detection by host antibodies that were elicited by prior infection or vaccination. Under this type of evolution, mutations that alter viral proteins to escape neutralization while retaining necessary viral functions are under positive selection. Thus, antigenic evolution causes the viral genome to continually fix nonsynonymous mutations in epitopes and can be detected as a high rate of adaptive evolution in genes encoding the targets of neutralizing antibodies.

To identify endemic human viruses that are evolving antigenically, we calculated rates of adaptation across the genomes of a wide diversity of viruses using an extension of the McDonald-Kreitman test.[26,28] This method divides an alignment of viral sequences into temporal windows and compares the isolates in each window with a fixed outgroup, which represents the historical genome sequence of that virus.[29,30] The number of adaptive mutations in each time window is calculated as an excess of fixed (or nearly fixed) nonsynonymous mutations above the neutral expectation. The rate of adaptation is then computed as the slope of the linear regression fitting adaptive mutations versus time. This temporal aspect means that recurrent fixations or selective sweeps over time will yield a high rate of adaptation, whereas a single adaptive fixation will not. Thus, this method is well suited toward our goal of detecting continuous adaptation, such as antigenic evolution.

However, because this method uses a fixed outgroup sequence, multiple mutations occurring within the same codon over time can give inaccurate results for a couple of reasons. First, whether a mutation is synonymous or nonsynonymous is determined by substituting that mutation into the outgroup sequence. This can result in a false assignment if a mutation within that codon has fixed previously. Additionally, repeated mutations at the same nucleotide position will be counted as only a single mutation because the method has no *knowledge*' of previous time windows. This flaw will cause a disproportionate underestimation of the rate of adaptation in viral proteins where many sites have sequentially fixed multiple mutations, as in rapidly evolving viruses such as influenza A/H3N2 (Figure 1A).

To address both these issues, we have modified the method to update the outgroup sequence each time a mutation fixes. Isolates in later time windows are, thus, *aware*' of any fixations that occurred in the same codon during previous time windows. This modification substantially affects the rate of adaptation in the influenza A/H3N2 hemagglutinin HA1 subunit, where 92 nucleotide sites have fixed nonsynonymous mutations since 1968, and 12 of these sites have seen multiple nonsynonymous fixations (Figure 1A). The asymptoting shape of the inferred accumulation of adaptive mutations in H3N2 HA1 reflects saturation where many adaptive mutations at later time windows occur at the same position as adaptive mutations that occurred in previous time windows (Figure 1C). In contrast to H3N2 HA1, there are zero nucleotide sites within the spike S1 subunit of seasonal coronavirus 229E that have fixed multiple nonsynonymous mutations. Fittingly, updating the outgroup sequence has little to no effect on the estimated rates of adaptation in the 229E receptor-binding subunit S1 (Figure 1F).

Influenza A/H3N2[9,20] and coronavirus 229E[7] are both known to undergo antigenic evolution through the fixation of mutations in their receptor-binding proteins. Because one goal of this manuscript is to quantitatively compare antigenic evolution between viruses by estimating rates of adaptive evolution in the receptor-binding proteins, we use the method that updates the outgroup sequence throughout this study, as it better captures the rate of a rapidly adapting protein, such as H3N2 HA1. However, it should be noted that the major findings and themes presented here do not depend on which version of the method is used—both methods identify the same subset of viruses as antigenically evolving, although the relative pace of this evolution is dependent on which method is used.

Estimation of the threshold of ongoing adaptive evolution

Antigenic evolution occurs when a virus fixes mutations at or near sites of antibody binding that abrogate those antibodies' abilities to neutralize the virus. For viruses that have been demonstrated to evolve antigenically, these escape mutations occur in the viral protein that mediates receptor binding, which is located on the virion's surface and is typically a primary target of neutralizing antibody binding. For instance, in influenza A/H3N2, antigenic evolution occurs largely in HA1, the receptor-binding subunit of hemagglutinin.[9,19–21] Similarly, for seasonal coronavirus 229E, escape mutations fix in S1, the receptor-binding subunit of spike.[7] Thus, we hypothesize that antigenic evolution results in a high rate of adaptation in the receptor-binding protein or subunit and that antigenically evolving viruses can be distinguished from antigenically stable ones by the rate of adaptation on the receptor-binding protein.

We calculated the rate of adaptive evolution in the receptor-binding protein for three viruses that are known to evolve antigenically—influenza viruses A/H3N2 and B/Yam[11] and coronavirus 229E[7]—as well as for three viruses that are known to be antigenically stable—measles,[16,17] influenza C/Yamagata,[18] and hepatitis A.[31,32] All of the antigenically evolving viruses have higher rates of adaptation than the antigenically stable viruses (Figure 2A), indicating that this method successfully differentiates between viruses that evolve antigenically and those that do not. We used these rates of adaptive evolution to estimate a threshold of antigenic evolution (i.e., a rate above which we predict the virus to be evolving antigenically) using logistic regression (see STAR Methods for more details). We estimated that the threshold of antigenic evolution is about 1.17×10^{-3}

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

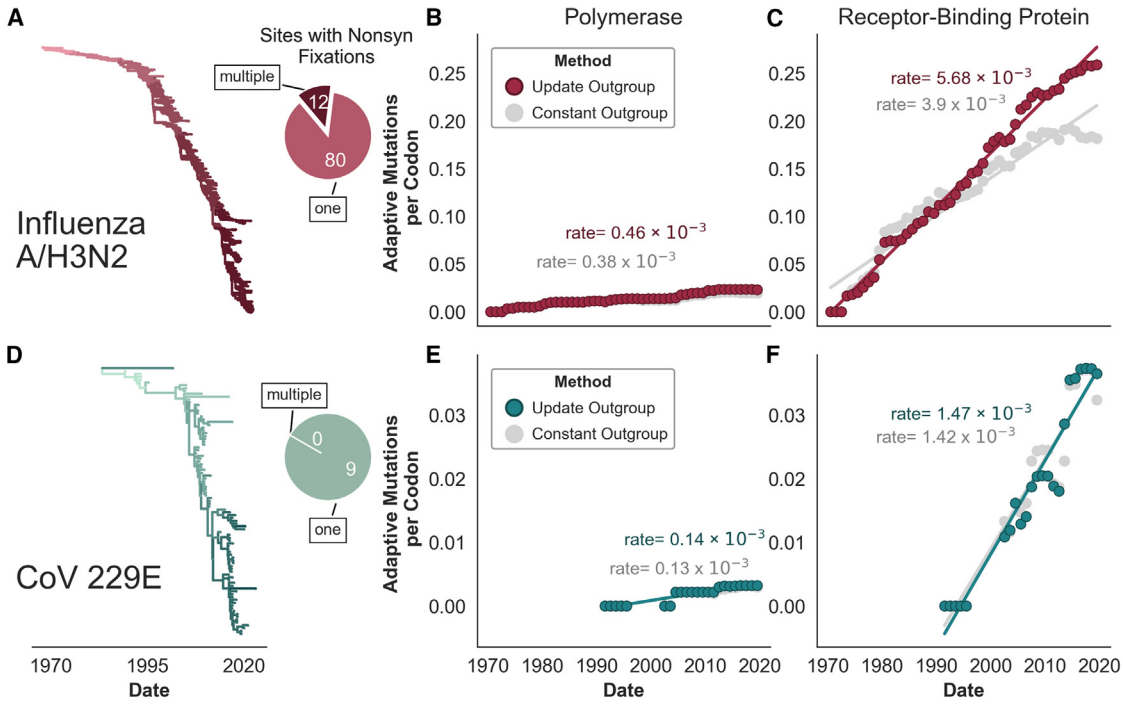


Figure 1. A McDonald-Kreitman-based method to estimate the rate of adaptation in antigenically evolving viruses

(A) Time-resolved phylogeny of 2,104 influenza A/H3N2 HA sequences sampled between 1968 and 2022 and colored by nonsynonymous mutation accumulation from the root, with darker reds symbolizing more mutations in the HA1 subunit. Within these samples, ninety-two nucleotide sites have completely fixed a nonsynonymous mutation, and the pie chart indicates that 12 of these nucleotide sites have fixed multiple nonsynonymous mutations during the past ~50 years.

(B) Accumulation of adaptive mutations (per codon) in polymerase PB1 as calculated by the McDonald-Kreitman-based method that updates the outgroup sequence at each fixation (dark red) or uses a constant outgroup sequence (gray). The rate of adaptation is the slope of the linear regression fitting these estimates.

(C) Estimated accumulation of adaptive mutations in HA1.

(D) Time-resolved phylogeny of 95 coronavirus 229E spike S1 sequences sampled between 1989 and 2022, colored as in (A). Pie chart indicates that, within these samples, nine nucleotide sites have completely fixed a nonsynonymous mutation, and zero nucleotide sites have fixed multiple nonsynonymous mutations.

(E and F) Accumulation of adaptive mutations, as in (B) and (C), within the coronavirus 229E (E) polymerase (RdRp) and (F) receptor-binding subunit S1.

mutations per codon per year in the receptor-binding protein. We view this threshold not as an absolute but, rather, as our best approximation of a rate that separates evolution with real biological meaning from noise in the estimates. Although we calculate this threshold based on proteins we know to be evolving antigenically, the threshold should be applicable to any viral protein undergoing any kind of ongoing adaptation that is occurring on a similar timescale to antigenic evolution.

The relative rates of adaptation estimated here are consistent with what is already known about the relative pace of antigenic evolution in these viruses. Bedford et al.[11] estimated that influenza A/H3N2 evolves antigenically 2–3 times faster than the influenza B viruses, and Eguia et al.[7] found that coronavirus 229E escapes neutralization at a rate similar to influenza B. Additionally, our estimated rates of adaptation in HA1 of the various influenza viruses reflect the frequencies at which these viruses deviate antigenically enough to warrant an update to the vaccine strain (Figure 2B). Influenza A/H3N2 exhibits the highest rate of adaptation (5.7×10^{-3} mutations per codon per year), followed by A/H1N1pdm (3.2×10^{-3} mutations per codon per year), then B/Vic (1.8×10^{-3} mutations per codon per year), and B/Yam (1.4×10^{-3} mutations per codon per year). Mirroring this, the A/H3N2 component of the vaccine has been updated 8 times (9 different strains) between 2012 and 2022, whereas the H1N1 strain was updated 4 times, the B/Vic component was updated 3 times, and B/Yam component was updated twice during this time period.[33] Components of the seasonal influenza vaccine are updated by the World Health Organization (WHO) Global Influenza Surveillance and Response System (GISRS) when the vaccine strain no longer induces sufficient protection against circulating viruses—a point that is typically defined by an 8-fold drop in titer in a hemagglutinin inhibition (HI) assay.[34] This suggests that the rate of adaptation in the receptor-binding domain can identify not only which viruses evolve antigenically but also the relative pace at which they do so and, thus, the expected duration of protection afforded by antibodies elicited by vaccination or infection.

Genome-wide appraisal of rapidly evolving viral proteins

We next sought to survey a wide diversity of endemic human viruses for evidence of ongoing adaptive evolution. We focus on viruses that have been endemic in humans for at least 12 years because we are interested in continued adaptive evolution that persists during the endemic phase (rather than initial host adaptation that occurs early in a pandemic) and because a short temporal spread of sampled sequences decreases the accuracy of

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

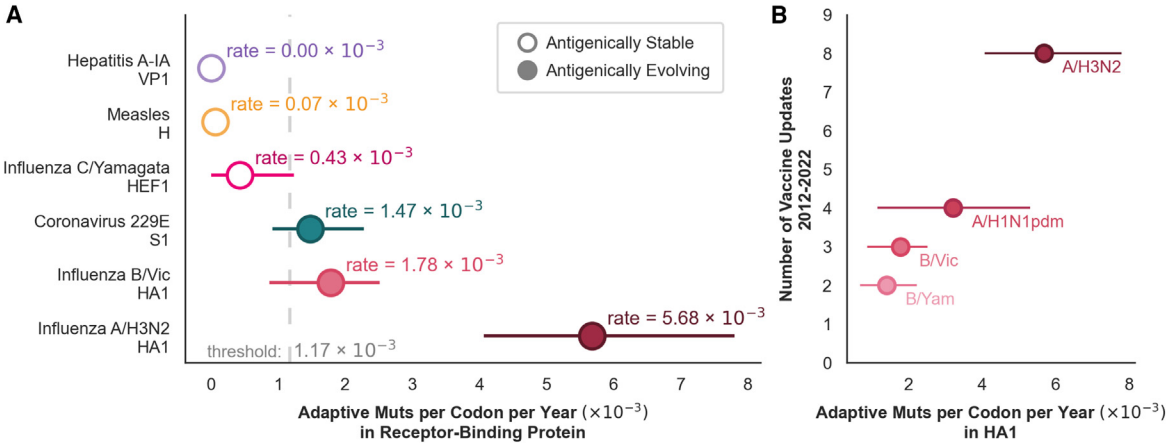


Figure 2. Rates of adaptation in the receptor-binding protein recapitulate known trends of antigenic evolution

(A) The rate of adaptation calculated in the receptor-binding protein is plotted for 3 antigenically stable viruses (solid circles) and 3 antigenically evolving viruses (open circles). The threshold of antigenic evolution is estimated by logistic regression. Error bars represent the 95% bootstrap percentiles.

(B) For each of the 4 influenza viruses that are included in the yearly flu vaccine, the rate of adaptation is compared with the number of times that the vaccine strain was updated between the 2012–2013 and 2022–2023 Northern hemisphere flu seasons.

the estimated rate of adaptation.[35] We analyze viruses at the subtype or genotype level and require that each viral subtype in the panel has at least 50 genomes spread over a minimum of 12 years (Figure S2 shows the temporal distribution of sequences for each virus). Although this *atlas*' aims to examine continuous adaptive evolution across a range of viral diversity, the panel of endemic viruses is far from comprehensive and largely limited by the availability of historical sequences, which, for many viruses, is not adequate to make accurate rate estimates. In total, we downloaded and curated sequence data for 28 human pathogenic viruses, which belong to 10 different viral families. This panel comprises both RNA and DNA viruses with a variety of modes of transmission, including respiratory transmission, fecal-oral transmission, vector-borne transmission, and transmission via blood or bodily fluids.

For each virus, we estimated the rate of adaptation in each gene of the genome. In total, we analyzed 239 viral genes, and 14 of them had rates of adaptation exceeding the threshold of antigenic evolution (Figure 3A). Of these 14 genes, 13 encode proteins located on the viral surface, with 10 of those being proteins that mediate host receptor binding (Figure 3B). The 3 surface proteins with rates of adaptation exceeding our threshold that are not classified as receptor binding are all neuraminadase (NA) genes of the influenza subtypes A/H3N2, A/H1N1pdm, and B/Vic. The influenza NA protein has been shown to bind host receptors in some influenza viruses[36,37] and to be a target of protective and neutralizing antibodies[38–40]; hence, it is possible that these high rates of adaptation in influenza NA are also reflective of evolution to escape antibody recognition. The non-surface protein that has a high rate of adaptation is norovirus p22, which antagonizes cellular protein trafficking[41] and has been previously reported to be under positive selection.[42]

In total, 10 of the 28 viruses in this panel had at least one gene we predict to be undergoing ongoing adaptation. These viruses include members of the orthomyxovirus (influenza A/H3N2, A/H1N1pdm, B/Vic, and B/Yam), paramyxovirus (RSV-A and RSV-B), coronavirus (229E and OC43-A), calicivirus (norovirus GII.4), and picornavirus (enterovirus D-68) families (Figure 3C). Although multiple orthomyxo-, corona-, and paramyxoviruses appear in this list, our results suggest that adaptive evolution is not necessarily a shared feature of related viruses. For instance, although influenza A/H3N2 has two adaptively evolving proteins, influenza C/Yamagata has none (Figures 3D and 3E). Similarly, 229E and NL63 are both alphacoronaviruses, but 229E has a high rate of adaptation in spike S1, whereas NL63 does not (Figures 3H and 3I).

Among the 10 viruses with at least one protein exceeding our threshold, we observe that the highest rates of adaptation genome wide are typically in the genes encoding the receptor-binding protein or subunit. The exception being influenza A/H1N1pdm and B/Yam, where the fastest rate is in NA, not HA1—although, as mentioned above, NA is sometimes involved in receptor binding. These results reveal that endemic viruses experience little-to-no ongoing adaptation throughout most of their genome and that continuous adaptive evolution is found almost solely in surface-exposed proteins, which are accessible to neutralizing antibodies. This suggests that evasion of antibody neutralization is a driving force in the ongoing adaptive evolution of many endemic viruses.

Identification of putative antigenically evolving viruses

With the expectation that antigenic evolution is detectable by a high rate of adaptation in the receptor-binding protein, we then directly compared rates between the receptor-binding proteins of 28 viruses (Figure 4). We also compared the rates of adaptation in the polymerase, which, in endemic viruses, we expect to be relatively conserved. We observe that although there is a little variation between the rates of adaptation in the polymerase, which range between 0.0 and 0.7×10^{-3} mutations per codon per year, there is a much larger spread of rates in the receptor-binding proteins of these viruses (ranging from 0.0 to 5.7×10^{-3} mutations per codon per year). Based on the estimated rates of adaptation that exceed the threshold for antigenic evolution we identified, we predict that 10 of the viruses in this panel evolve antigenically.

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

CellPress
OPEN ACCESS

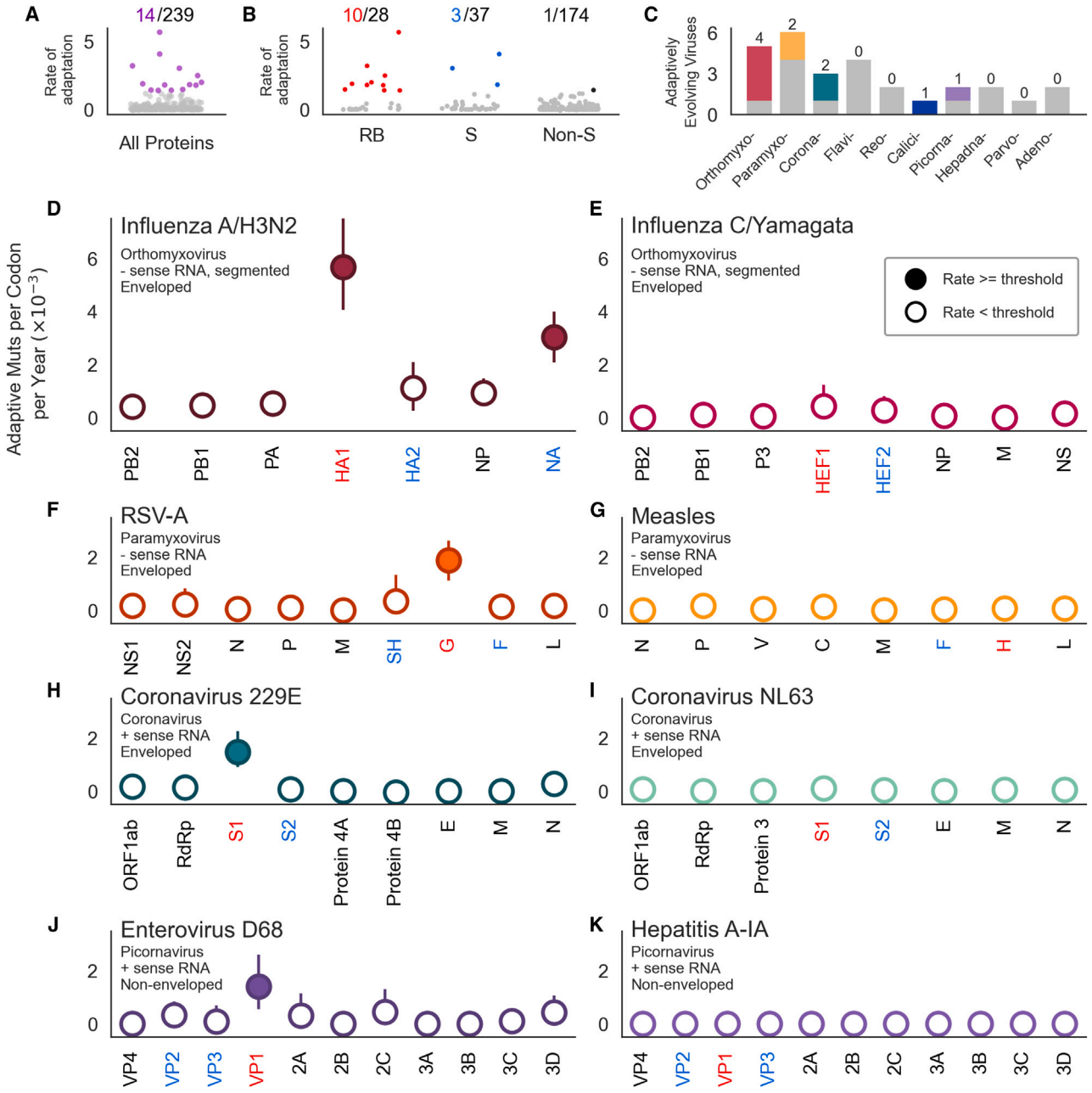


Figure 3. Across 28 viral genomes, the highest rates of adaptation are found in surface-located receptor-binding proteins

(A) The rate of adaptation for all 239 viral genes. Fourteen genes (in purple) have rates of adaptation above our threshold of antigenic evolution. Genes with rates of adaptation below the threshold are in gray.

(B) The rate of adaptation within all 28 receptor-binding proteins (RB, left), 37 other proteins located on the viral surface (S, center), and 174 non-surface proteins (non-S, right). Ten receptor-binding proteins (red), 3 other surface-located proteins (blue), and 1 non-surface protein (black) exceed our threshold. Genes with rates below the threshold are in gray.

(C) Number of viruses per viral family that have at least one gene exceeding the threshold are shown in color. The number of viruses in these families that had no high rates of adaptation throughout their entire genome is in gray.

(D–K) Rates of adaptation were calculated for each gene, subunit, or coding region indicated along the x-axis, and ordered by genomic position (or segment number, for segmented viruses). Receptor-binding proteins are labeled in red, other surface-exposed proteins are in blue, and non-surface-located proteins are in black. Filled circles indicate genes with rates exceeding the threshold. Each row shows two viruses from the same viral family, one that contains at least one adaptively evolving gene (left) and one that does not (right). Error bars indicate the 95% bootstrap percentiles from 100 bootstrapped data sets.

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

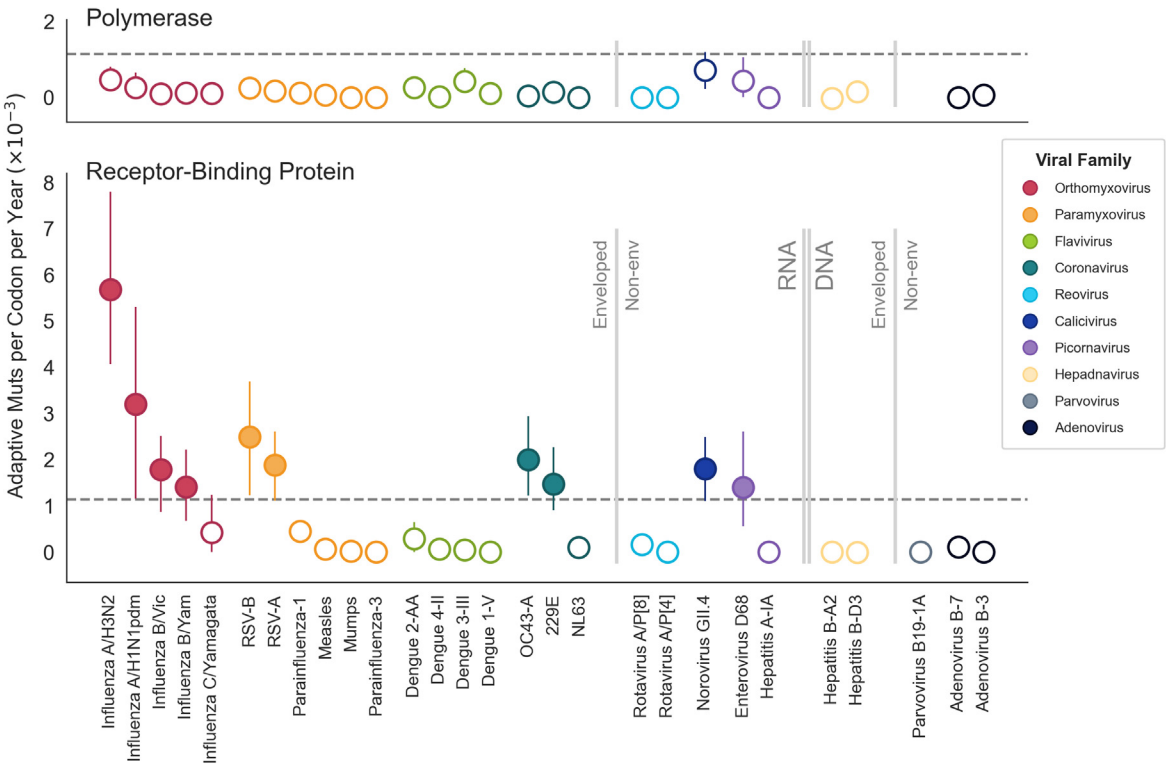


Figure 4. Comparison of rates of predicted antigenic evolution across a wide diversity of human pathogenic viruses

Rates of adaptive evolution in the polymerase (top) and receptor-binding protein (bottom) for 28 human pathogenic viruses. The receptor-binding and polymerase genes for each virus are listed in Table S1. Error bars indicate the 95% bootstrap percentiles from 100 bootstrapped data sets. The threshold of antigenic evolution (as determined in Figure 2) is marked by the dotted line; the rates falling above this line are shown by solid markers, and the rates below the threshold are open circles. Viruses are grouped and colored by viral family and arranged within viral family in descending order of the receptor-binding rate. Viral families are ordered by genome type, with RNA viruses shown in brighter colors and DNA viruses in gray tones. Vertical dividers further delineate enveloped from non-enveloped viruses.

Influenza A/H3N2 has, by far, the fastest rate of antigenic evolution, followed by A/H1N1pdm, whereas the other 8 putative antigenically evolving viruses all have rates in roughly the same range, around 1.5 to 2×10^{-3} adaptive mutations per codon per year. In descending order of the estimated rate of antigenic evolution, these viruses are as follows: RSV-B, coronavirus OC43-A, RSV-A, norovirus GII.4, influenza B/Vic, coronavirus 229E, influenza B/Yam, and enterovirus D-68.

All of these 10 viruses have RNA genomes, although both positive- and negative-sense genomes and both enveloped and non-enveloped viruses appear capable of antigenic evolution. All of these viruses transmit via a respiratory route except norovirus, which uses fecal-oral transmission. We observe that multiple coronaviruses, RSV viruses, and influenza viruses evolve antigenically, suggesting that these types of viruses might have a higher propensity for this type of evolution. However, in each of these cases, we also find that at least one other member of the same viral family does not evolve antigenically, indicating that relatedness at the level of viral family is not an absolute predictor of antigenic evolution. Overall, the examination of antigenic evolution presented here suggests that selection to evade antibody recognition is widespread, although certainly not ubiquitous, among endemic RNA viruses and that although certain types of viruses may have a higher propensity for this type of evolution, even closely related viruses can differ in this regard.

A comparison of rates of adaptation in the receptor-binding proteins of all the viruses in this panel as well as rates across the genome of each of these viruses can be viewed interactively at blab.github.io/atlas-of-viral-adaptation/ (see example screenshots in Figure S1). This interactive website shows the rates per codon per year (as reported in this manuscript) as well as per gene per year, allows the viruses to be ordered by rate rather than by viral family, and displays rates calculated both by the constant outgroup and updated outgroup methods.

Comparison of rates of evolution between endemic viruses and SARS-CoV-2

An obvious question is where the evolution of SARS-CoV-2 falls with respect to these other viruses. Since the beginning of the SARS-CoV-2 pandemic, we have seen a period of many co-circulating variants that contained adaptive mutations, which was followed by a single fixation event where Omicron swept and a subsequent period where many competing Omicron lineages are co-circulating, with repeated near sweeps of derived lineages such as BA.5, BQ.1, and XBB.1.5 that are supplanted before reaching fixation. Because the McDonald-Kreitman-based rate estimation we have used thus far considers only fixed or nearly

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

CellPress
OPEN ACCESS

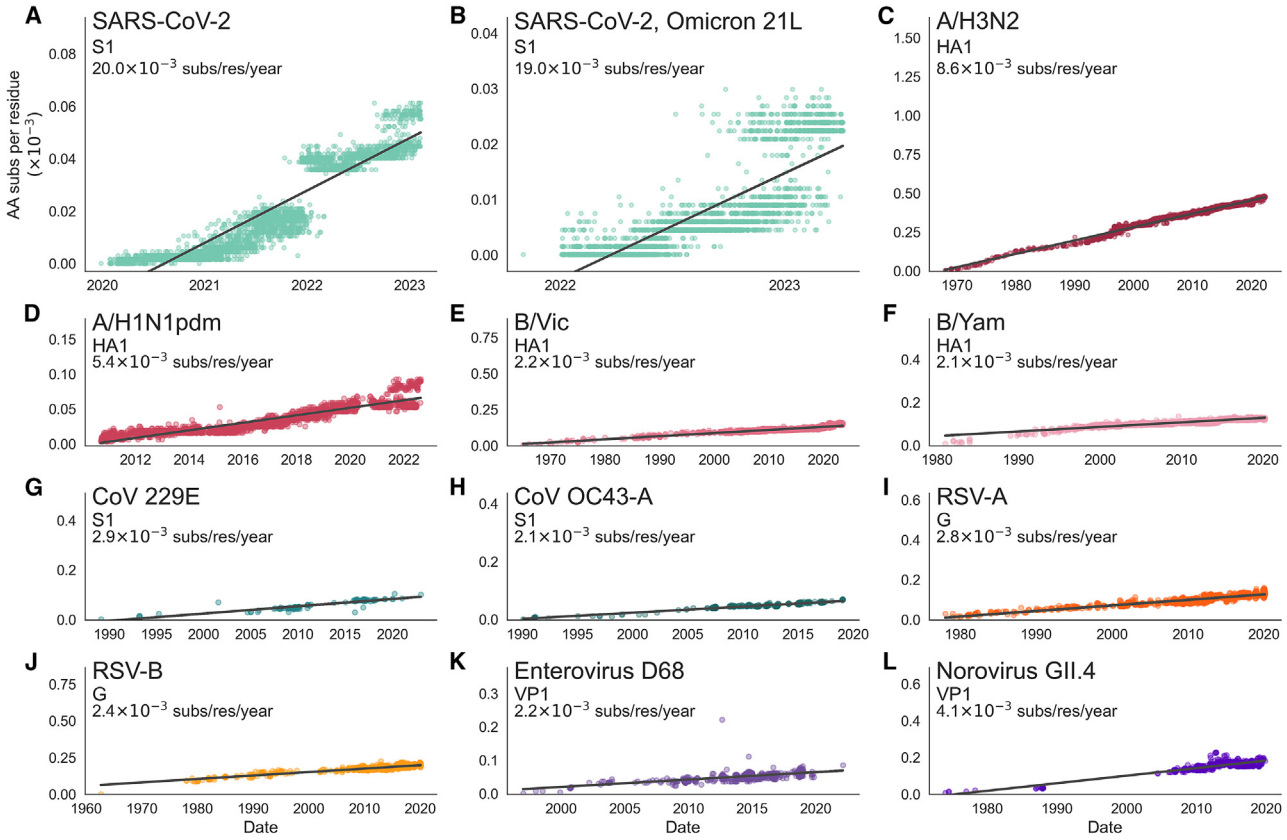


Figure 5. Rates of amino acid substitution in the receptor-binding protein of SARS-CoV-2 and 10 antigenically evolving endemic viruses

The rate of amino acid substitution in the receptor-binding protein of (A) SARS-CoV-2, (B) SARS-CoV-2 Omicron clade 21L, (C) influenza A/H3N2, (D) influenza A/H1N1pdm, (E) influenza B/Vic, (F) influenza B/Yam, (G) coronavirus 229E, (H) coronavirus OC43-A, (I) RSV-A, (J) RSV-B, (K) enterovirus D68, and (L) norovirus GII.4. The receptor-binding protein, or subunit is labeled below the virus name. Rates are computed as the slope of a linear regression fitting a comparison of amino acid substitutions versus time and are found using a phylogeny. Each tip on the tree is plotted by its sampling date and the number of amino acid substitutions that accumulated between the root and the tip (normalized by the length of the coding region, in residues). Aspect ratios in each panel are fixed so that regression slopes are visually comparable across panels.

fixed mutations to be potentially adaptive, this method is ill suited to analyzing SARS-CoV-2 evolution so far. Essentially, the calculated rate will just reflect the high number of mutations on the long branch leading to Omicron that fixed when Omicron swept. Additionally, this method can be noisy over short time periods, where small numbers of fixations can have an outsized effect on the rate. It is for this reason that we limited our panel to viruses that have been endemic for several years, with influenza A/H1N1pdm having the narrowest span of human circulation (12 years).

In lieu of calculating a rate of adaptation for SARS-CoV-2, we instead do a much simpler comparison of the rates of amino acid substitution in the receptor-binding proteins between SARS-CoV-2 and the 10 viruses we predict to be evolving antigenically (Figure 5). We find that SARS-CoV-2 accumulates roughly 20×10^{-3} amino acid substitutions per residue per year in S1. This is 2–2.5× faster than the accumulation of amino acid substitutions in influenza A/H3N2 HA1 and 7–10× faster than in the S1 subunit of seasonal coronaviruses 229E and OC43. Importantly, the rate of amino acid substitution among all SARS-CoV-2 viruses is not solely driven by the fixation of 12 S1 substitutions when Omicron swept; in fact, the rate we observe between all SARS-CoV-2 viruses (Figure 5A) is roughly the same as the rate among just the currently predominant clade of Omicron 21L and its descendants (Figure 5B), corresponding to lineage BA.2 and derived lineages such as BQ.1 and XBB.

In Figure 5, we plot the number of amino acid changes per residue in the receptor-binding protein that each tip has compared with the root. This simpler analysis does not consider the fixation of particular mutations, nor does it make any attempt to account for substitutions under selection versus those that are present due to chance or hitch-hiking. However, we find that this analysis reflects the general relationships between rates of antigenic evolution of different viruses that we present in Figure 4. Figure 6 lists the rate of amino acid substitution and the rate of adaptation in the receptor-binding protein of each virus. A ratio of the rates in Figures 4 and 5 indicates that in most antigenically evolving endemic viruses, between ~60% and 100% of amino acid substitutions in the receptor-binding protein are adaptive. For instance, we estimate that influenza A/H3N2 evolves antigenically at 5.7×10^{-3} adaptive mutations per codon per year, which is ~66% of the 8.6×10^{-3} amino acid substitutions per residue it accumulates each year. Of the 10 antigenically evolving endemic viruses, the lowest proportion of amino acid substitutions that are adaptive is 44% in norovirus GII.4 VP1.

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

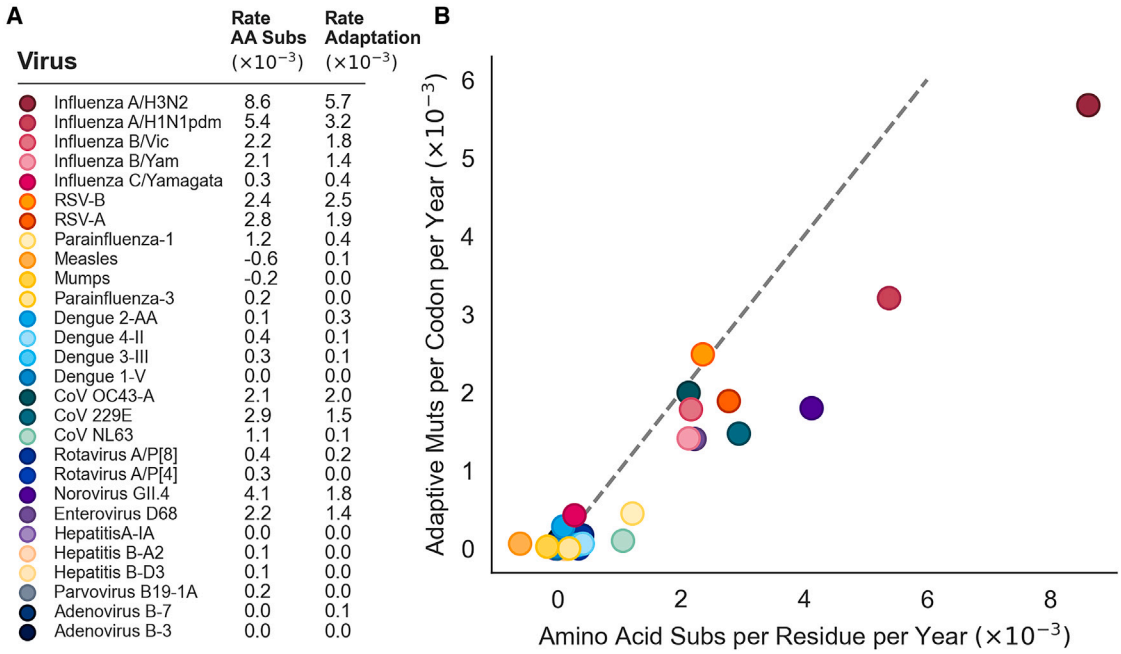


Figure 6. Comparison of rates of amino acid substitution to rates of adaptation

(A) The rate of amino acid substitution ($\times 10^{-3}$) and rate of adaptive evolution ($\times 10^{-3}$) is listed for each of the 28 viruses in the panel.

(B) Rate of amino acid substitution is plotted against rate of adaptive evolution for each virus, with color corresponding to the panel A. The dashed gray line is drawn at X = Y to indicate the point where all amino acid substitutions are inferred to be adaptive.

DISCUSSION

In search of antigenically evolving viruses, we use genomic sequences to analyze adaptive evolution across a panel of 28 viruses and present the results here and as a website with interactive plots and phylogenies at blab.github.io/atlas-of-viral-adaptation. We find that antigenic evolution is not uncommon among endemic human viruses, with ten viruses spanning five viral families meeting our criteria for predicting antigenic evolution. Particularly, this mode of evolution seems prevalent among endemic viruses that have RNA genomes (Figure 4). However, the true proportion of endemic viruses that evolve antigenically is hard to estimate because the panel of viruses analyzed here is far from a comprehensive list of human endemic viruses and is biased toward well-studied viruses, which are not necessarily the most common pathogens. We selected viruses to include in the panel based on the following criteria: (1) virus has been endemic for at least 12 years, (2) the genome is under 50 kb, and (3) there are at least 50 high-quality genomes available spanning at least 12 years. For many endemic viruses, the limiting factor is a dearth of historical sequences predating the mid-2000's. However, the COVID-19 pandemic has spurred an increased interest in monitoring and sequencing human pathogens, and if this trend continues, it is likely that there will be enough longitudinal data to add many more viruses to this panel in the years to come.

By employing a quantitative method, we are able to compare the pace of adaptive evolution between genes in a genome as well as between viruses. Comparisons within genomes reveal that surface proteins are consistently the fasting-evolving viral proteins (Figure 3). Comparisons between viruses show that influenza A/H3N2 is especially striking its pace of antigenic evolution, which is roughly 2–3 times faster than all other viruses we predict to be evolving antigenically (Figure 4). The observation that many viruses accumulate adaptive mutations in their receptor-binding protein at a rate of roughly 1.5 to 2.0×10^{-3} mutations per codon per year suggests that this might be a lower bound on the rate that is sufficient to generate antigenic novelty fast enough that a virus can persist in a mostly immune population. However, it should be stressed that although the rate of adaptation is similar between these eight viruses, it is not clear whether the pace of relevant phenotypic change is also similar between them. For instance, it may be that some viruses need, say, two adaptive mutations on average to successfully escape prior immunity, whereas other viruses need only one.

Whether a virus evolves antigenically, and the pace at which it does so, is likely a function of many factors, including mutation rate, mutational tolerance of surface proteins,[43] positioning and co-dominance of epitopes,[17] viral transmission dynamics, and existing population immunity. Our estimates of rates of antigenic evolution do not allow us to disentangle which factors are contributing most to the evolution of endemic viral proteins. However, the questions of why closely related viruses (such as coronaviruses 229E and NL63) differ in their propensity to evolve evasion of antibody detection and, relatedly, what the minimal necessary information is to predict this type of evolution in an emerging virus are interesting and open questions.

In this study, we have focused on continuous antigenic evolution within viral lineages over the past ~50 years. It is important to note that this is a very particular type of evolution in which antigenic variation is selected for repeatedly, leading to selective sweeps within a single lineage. However, this does not

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

necessarily mean that viruses that are antigenically stable on a ~50-year time scale have not undergone some form of antigenic evolution in the past or at another timescale. For instance, influenza C viruses and dengue viruses both exist as several antigenically distinct lineages, and although ongoing antigenic evolution is not occurring *within* these lineages, the establishment of antigenically distinct lineages was likely a result of selection in the past. It is possible that some viruses are more prone to fracturing into several antigenically distinct, co-circulating lineages rather than undergoing perpetual antigenic evolution within a lineage.

Relatedly, it is important to note that this method looks for fixations and near fixations, with the idea that positively selected mutations will sweep through the population. This means that mutations that fix within a clade, but not the entire population, will not be considered potentially adaptive and, thus, that this method is sensitive to how lineages are designated. For instance, if all influenza B viruses were analyzed together, rather than as separate B/Vic and B/Yam lineages, there would be no signal of adaptation in the HA surface protein. In some cases, it can be difficult to define what constitutes two distinct lineages versus two clades of the same lineage. In our analyses, we have divided each viral species into the lineage or genotype classifications used by the field of literature for that virus. As noted above, if a virus is composed of multiple geographically—or ecologically—distinct lineages, sweeping adaptive mutations occurring on one or both lineages will be obscured in our analysis, as they will appear to be persistent polymorphisms. However, in such a case, we would still expect to see a high absolute rate of amino acid substitution, as separate lineages would each accumulate substitutions. Here, all of the viruses with low rates of adaptation also have low rates of amino acid substitution (Figure 6), indicating that we are not missing adaptively evolving viruses due to population structure.

Implications for the ongoing evolution of SARS-CoV-2

In SARS-CoV-2 spike S1, we observe a rate of amino acid substitution that is roughly 2–2.5× the rate in influenza A/H3N2 HA1, the prototypical example of rapid antigenic evolution. There is an open question of whether SARS-CoV-2 can sustain such high rates of evolution in the years to come. To address this question, we can retroactively observe how the evolution of other viruses has changed between the early pandemic and the ensuing endemic years. In this manuscript, we have analyzed the evolution of influenza viruses A/H3N2 and A/H1N1pdm over time between their respective introductions in 1968 and 2009 and today. We do not see any evidence that the rate of amino acid substitution (Figures 5C and 5D) or the rate of adaptive evolution (Figure 1C) in HA1 is flagging in these antigenically evolving influenza viruses. By extension, this suggests that we may continue to see rapid evolution in the S1 subunit of SARS-CoV-2. This fits with our observation that the rate of amino acid substitution in the S1 of Omicron clade 21L viruses circulating in 2022 and 2023 is roughly the same as this rate over the entire pandemic (Figures 5A and 5B), suggesting that, so far, S1 evolution has not slowed throughout the course of the pandemic. Additionally, this suggests that the overall rate of adaptation is not a simple proxy for initial post-spillover host adaptation vs. longer-term continued antigenic drift. Future work on SARS-CoV-2 examining phenotypic effects of spike mutations on ACE2 binding vs. immune escape and their adaptive impacts could distinguish crossover from initial host adaptation to later continued antigenic drift.

Although overall we expect that SARS-CoV-2 will continue to evolve at appreciably faster rates than seasonal influenza or coronaviruses, it is unclear whether this evolution will be somewhat slowed by the build-up of increasingly complex immune histories toward this virus. At this point, it is also difficult to predict whether the emergence of a highly fit and highly divergent variant (Omicron) was a one-time event or whether other similar lineages will emerge in the future and continue to be a feature of SARS-CoV-2 evolution.

The fact that SARS-CoV-2 is able to evolve antigenically has become readily apparent over the three years since the beginning of the pandemic.[13–15,25] However, in early 2020, at the beginning of the COVID-19 pandemic, it was not known whether related coronaviruses evolve antigenically, and thus, it was difficult to speculate whether SARS-CoV-2 would evolve in this way. We believe that this issue reveals how little is known about the antigenic evolution of many of the viruses that commonly infect us. We believe that a better understanding of the broad diversity of endemic viruses will not only better prepare us for future pandemics but also inform our current efforts to design vaccines and therapeutics against these viruses. To this end, we have compiled this atlas of viral adaptive evolution to quantitatively compare evolution across a wide range of endemic viruses.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Input data for the estimation of rate of adaptation
  - Sequence data
  - Nextstrain builds to generate alignments and trees
  - SARS-CoV-2 Nextstrain phylogenies
  - Rate of adaptation, with a fixed outgroup
  - Rate of adaptation, with an updated outgroup
  - Estimation of threshold, using logistic regression
  - Rate of amino acid substitution
  - Genes analyzed for each virus
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2023.09.012.

ACKNOWLEDGMENTS

We gratefully acknowledge the authors who generated and submitted the viral sequences deposited in the NCBI GenBank and ViPR/BV-BRC databases, on which the analyses in this manuscript are based. We also appreciatively acknowledge the authors, originating and submitting laboratories of the SARS-CoV-2 genetic sequences and metadata made available through GISAID Initiative, which supplied the data for the SARS-CoV-2 analyses in

this manuscript. We have included an acknowledgements table for NCBI GenBank, ViPR/BVBRC, and GISAID sequences in supplemental information. We also thank Allison Li for putting together the norovirus Nextstrain build. Finally, we thank Dr. Harmit Malik for the idea that sparked this project; Dr. John Huddleston for giving it a name; and Cassia Wagner, Dr. Cécile Tran Kiem, and Dr. John Huddleston for feedback on the manuscript. T.B. is a Howard Hughes Medical Institute investigator. K.K. is also supported by Howard Hughes Medical Institute. The graphical abstract was created with BioRender.com.

AUTHOR CONTRIBUTIONS

Conceptualization, K.E.K. and T.B.; methodology, K.E.K.; software, K.E.K.; formal analysis, K.E.K.; investigation, K.E.K.; writing – original draft, K.E.K.; writing – review and editing, K.E.K. and T.B.; visualization, K.E.K.; funding acquisition, K.E.K. and T.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: May 22, 2023
Revised: August 25, 2023
Accepted: September 28, 2023
Published: October 25, 2023

REFERENCES

1. Meyer, A.G., Spielman, S.J., Bedford, T., and Wilke, C.O. (2015). Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evol. *1*, vev006. https://doi.org/10.1093/ve/vev006.
2. Su, Y.C.F., Bahl, J., Joseph, U., Butt, K.M., Peck, H.A., Koay, E.S.C., Oon, L.L.E., Barr, I.G., Vijaykrishna, D., and Smith, G.J.D. (2015). Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. Nat. Commun. *6*, 7952. https://doi.org/10.1038/ncomms8952.
3. Kistler, K.E., Huddleston, J., and Bedford, T. (2022). Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host Microbe *30*, 545–555.e4. https://doi.org/10.1016/j.chom.2022.03.018.
4. Tao, K., Tzou, P.L., Nouhin, J., Gupta, R.K., de Oliveira, T., Kosakovsky Pond, S.L., Fera, D., and Shafer, R.W. (2021). The biological and clinical significance of emerging SARS-CoV-2 variants. Nat. Rev. Genet. *22*, 757–773. https://doi.org/10.1038/s41576-021-00408-x.
5. Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., and Fouchier, R.A.M. (2004). Mapping the antigenic and genetic evolution of influenza virus. Science *305*, 371–376. https://doi.org/10.1126/science.1097211.
6. Plotkin, J.B., Dushoff, J., and Levin, S.A. (2002). Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc. Natl. Acad. Sci. USA *99*, 6263–6268. https://doi.org/10.1073/pnas.082110799.
7. Eguia, R.T., Crawford, K.H.D., Stevens-Ayers, T., Kelnhofer-Millevolte, L., Greninger, A.L., Englund, J.A., Boeckh, M.J., and Bloom, J.D. (2021). A human coronavirus evolves antigenically to escape antibody immunity. PLoS Pathog. *17*, e1009453. https://doi.org/10.1371/journal.ppat.1009453.
8. Carrat, F., and Flahault, A. (2007). Influenza vaccine: the challenge of antigenic drift. Vaccine *25*, 6852–6862. https://doi.org/10.1016/j.vaccine.2007.07.027.
9. Koel, B.F., Burke, D.F., Bestebroer, T.M., van der Vliet, S., Zondag, G.C.M., Vervaet, G., Skepner, E., Lewis, N.S., Spronken, M.I.J., Russell, C.A., et al. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science *342*, 976–979. https://doi.org/10.1126/science.1244730.
10. Li, C., Hatta, M., Burke, D.F., Ping, J., Zhang, Y., Ozawa, M., Taft, A.S., Das, S.C., Hanson, A.P., Song, J., et al. (2016). Selection of antigenically advanced variants of seasonal influenza viruses. Nat. Microbiol. *1*, 16058. https://doi.org/10.1038/nmicrobiol.2016.58.
11. Bedford, T., Suchard, M.A., Lemey, P., Dudas, G., Gregory, V., Hay, A.J., McCauley, J.W., Russell, C.A., Smith, D.J., and Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. eLife *3*, e01914. https://doi.org/10.7554/eLife.01914.
12. Harvey, W.T., Benton, D.J., Gregory, V., Hall, J.P.J., Daniels, R.S., Bedford, T., Haydon, D.T., Hay, A.J., McCauley, J.W., and Reeve, R. (2016). Identification of low- and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A(H1N1) viruses. PLoS Pathog. *12*, e1005526. https://doi.org/10.1371/journal.ppat.1005526.
13. Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P.D., et al. (2021). Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. Nature *593*, 130–135.
14. Ai, J., Wang, X., He, X., Zhao, X., Zhang, Y., Jiang, Y., Li, M., Cui, Y., Chen, Y., Qiao, R., et al. (2022). Antibody evasion of SARS-CoV-2 Omicron BA.1, BA.1.1, BA.2, and BA.3 sub-lineages. Cell Host Microbe *30*, 1077–1083.e4. https://doi.org/10.1016/j.chom.2022.05.001.
15. Dejnirattisai, W., Huo, J., Zhou, D., Zahradník, J., Supasa, P., Liu, C., Duyvesteyn, H.M.E., Ginn, H.M., Mentzer, A.J., Tuekprakhon, A., et al. (2022). SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. Cell *185*, 467–484.e15. https://doi.org/10.1016/j.cell.2021.12.046.
16. Tahara, M., Ito, Y., Brindley, M.A., Ma, X., He, J., Xu, S., Fukuhara, H., Sakai, K., Komase, K., Rota, P.A., et al. (2013). Functional and structural characterization of neutralizing epitopes of measles virus hemagglutinin protein. J. Virol. *87*, 666–675. https://doi.org/10.1128/JVI.02033-12.
17. Muñoz-Alía, M.Á., Nace, R.A., Zhang, L., and Russell, S.J. (2021). Serotypic evolution of measles virus is constrained by multiple co-dominant B cell epitopes on its surface glycoproteins. Cell Rep. Med. *2*, 100225. https://doi.org/10.1016/j.xcrm.2021.100225.
18. Matsuzaki, Y., Sugawara, K., Abiko, C., Ikeda, T., Aoki, Y., Mizuta, K., Katsushima, N., Katsushima, F., Katsushima, Y., Itagaki, T., et al. (2014). Epidemiological information regarding the periodic epidemics of influenza C virus in Japan (1996–2013) and the seroprevalence of antibodies to different antigenic groups. J. Clin. Virol. *61*, 87–93. https://doi.org/10.1016/j.jcv.2014.06.017.
19. Wiley, D.C., Wilson, I.A., and Skehel, J.J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature *289*, 373–378. https://doi.org/10.1038/289373a0.
20. Underwood, P.A. (1982). Mapping of antigenic changes in the haemagglutinin of Hong Kong influenza (H3N2) strains using a large panel of monoclonal antibodies. J. Gen. Virol. *62*, 153–169. https://doi.org/10.1099/0022-1317-62-1-153.
21. Chambers, B.S., Parkhouse, K., Ross, T.M., Alby, K., and Hensley, S.E. (2015). Identification of hemagglutinin residues responsible for H3N2 antigenic drift during the 2014–2015 influenza season. Cell Rep. *12*, 1–6. https://doi.org/10.1016/j.celrep.2015.06.005.
22. Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J.C., Muecksch, F., Rutkowska, M., Hoffmann, H.-H., Michailidis, E., et al. (2020). Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife *9*, e61312. https://doi.org/10.7554/eLife.61312.
23. Liu, Z., VanBlargan, L.A., Bloyet, L.-M., Rothlauf, P.W., Chen, R.E., Stumpf, S., Zhao, H., Errico, J.M., Theel, E.S., Liebeskind, M.J., et al. (2021). Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. Cell Host Microbe *29*, 477–488.e4. https://doi.org/10.1016/j.chom.2021.01.014.

24. Kuzmina, A., Khalaila, Y., Voloshin, O., Keren-Naus, A., Boehm-Cohen, L., Raviv, Y., Shemer-Avni, Y., Rosenberg, E., and Taube, R. (2021). SARS-CoV-2 spike variants exhibit differential infectivity and neutralization resistance to convalescent or post-vaccination sera. Cell Host Microbe *29*, 522–528.e2. https://doi.org/10.1016/j.chom.2021.03.008.

25. Greaney, A.J., Starr, T.N., Barnes, C.O., Weisblum, Y., Schmidt, F., Caskey, M., Gaebler, C., Cho, A., Agudelo, M., Finkin, S., et al. (2021). Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. Nat. Commun. *12*, 4196. https://doi.org/10.1038/s41467-021-24435-8.

26. McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654. https://doi.org/10.1038/351652a0.

27. Smith, N.G.C., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–1024. https://doi.org/10.1038/4151022a.

28. Williamson, S. (2003). Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol. Biol. Evol. *20*, 1318–1325. https://doi.org/10.1093/molbev/msg144.

29. Bhatt, S., Katzourakis, A., and Pybus, O.G. (2010). Detecting natural selection in RNA virus populations using sequence summary statistics. Infect. Genet. Evol. *10*, 421–430. https://doi.org/10.1016/j.meegid.2009.06.001.

30. Bhatt, S., Holmes, E.C., and Pybus, O.G. (2011). The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. *28*, 2443–2451. https://doi.org/10.1093/molbev/msr044.

31. Gust, I.D., Lehmann, N.I., Crowe, S., McCrorie, M., Locarnini, S.A., and Lucas, C.R. (1985). The origin of the HM175 strain of hepatitis A virus. J. Infect. Dis. *151*, 365–367. https://doi.org/10.1093/infdis/151.2.365.

32. Armstrong, M.E., Giesa, P.A., Davide, J.P., Redner, F., Waterbury, J.A., Rhoad, A.E., Keys, R.D., Provost, P.J., and Lewis, J.A. (1993). Development of the formalin-inactivated hepatitis A vaccine, VAQTA from the live attenuated virus strain CR326F. J. Hepatol. *18*, S20–S26. https://doi.org/10.1016/S0168-8278(05)80373-3.

33. World Health Organization (WHO). Recommended composition of influenza virus vaccines for use in the 2012-2013 northern hemisphere influenza season. https://web.archive.org/web/20130301015922/http://www.who.int/influenza/vaccines/virus/recommendations/2012_13_north/en/index.html. Retrieved 18 Oct 2023.

34. Xie, H., Wan, X.-F., Ye, Z., Plant, E.P., Zhao, Y., Xu, Y., Li, X., Finch, C., Zhao, N., Kawano, T., et al. (2015). H3N2 mismatch of 2014–15 northern hemisphere influenza vaccines and head-to-head comparison between human and ferret antisera derived antigenic maps. Sci. Rep. *5*, 15279. https://doi.org/10.1038/srep15279.

35. Kistler, K.E., and Bedford, T. (2021). Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. eLife *10*, e64509. https://doi.org/10.7554/eLife.64509.

36. Zhu, X., McBride, R., Nycholat, C.M., Yu, W., Paulson, J.C., and Wilson, I.A. (2012). Influenza virus neuraminidases with reduced enzymatic activity that avidly bind sialic acid receptors. J. Virol. *86*, 13371–13383. https://doi.org/10.1128/JVI.01426-12.

37. Hooper, K.A., and Bloom, J.D. (2013). A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. J. Virol. *87*, 12531–12540. https://doi.org/10.1128/JVI.01889-13.

38. Chen, Y.-Q., Wohlbold, T.J., Zheng, N.-Y., Huang, M., Huang, Y., Neu, K.E., Lee, J., Wan, H., Rojas, K.T., Kirkpatrick, E., et al. (2018). Influenza infection in humans induces broadly cross-reactive and protective neuraminidase-reactive antibodies. Cell *173*, 417–429.e10. https://doi.org/10.1016/j.cell.2018.03.030.

39. Gilbert, P.B., Fong, Y., Juraska, M., Carpp, L.N., Monto, A.S., Martin, E.T., and Petrie, J.G. (2019). HAI and NAI titer correlates of inactivated and live attenuated influenza vaccine efficacy. BMC Infect. Dis. *19*, 453. https://doi.org/10.1186/s12879-019-4049-5.

40. Stadlbauer, D., Zhu, X., McMahon, M., Turner, J.S., Wohlbold, T.J., Schmitz, A.J., Strohmeier, S., Yu, W., Nachbagauer, R., Mudd, P.A., et al. (2019). Broadly protective human antibodies that target the active site of influenza virus neuraminidase. Science *366*, 499–504. https://doi.org/10.1126/science.aay0678.

41. Sharp, T.M., Guix, S., Katayama, K., Crawford, S.E., and Estes, M.K. (2010). Inhibition of cellular protein secretion by Norwalk virus nonstructural protein p22 requires a mimic of an endoplasmic reticulum export signal. PLoS One *5*, e13130. https://doi.org/10.1371/journal.pone.0013130.

42. Cotten, M., Petrova, V., Phan, M.V.T., Rabaa, M.A., Watson, S.J., Ong, S.H., Kellam, P., and Baker, S. (2014). Deep sequencing of Norovirus genomes defines evolutionary patterns in an urban tropical setting. J. Virol. *88*, 11056–11069. https://doi.org/10.1128/JVI.01333-14.

43. Thyagarajan, B., and Bloom, J.D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLife *3*, e03300. https://doi.org/10.7554/eLife.03300.

44. Huddleston, J., Hadfield, J., Sibley, T.R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R.A., Hodcroft, E.B. (2021). Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. J. Open Source Softw. 6. https://doi.org/10.21105/joss.02906.

45. Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. *4*, vex042. https://doi.org/10.1093/ve/vex042.

46. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

47. Aksamentov, I., Roemer, C., Hodcroft, E., and Neher, R. (2001). Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. *6*, 3773. https://doi.org/10.21105/joss.03773.

48. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.

49. Neher, R.A., and Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics *31*, 3546–3548. https://doi.org/10.1093/bioinformatics/btv381.

50. Moncla, L.H., Black, A., DeBolt, C., Lang, M., Graff, N.R., Pérez-Osorio, A.C., Müller, N.F., Haselow, D., Lindquist, S., and Bedford, T. (2021). Repeated introductions and intensive community transmission fueled a mumps virus outbreak in Washington State. eLife *10*, e66448. https://doi.org/10.7554/eLife.66448.

51. Bell, S.M., Katzelnick, L., and Bedford, T. (2019). Dengue genetic divergence generates within-serotype antigenic variation, but serotypes dominate evolutionary dynamics. eLife *8*, e42496. https://doi.org/10.7554/eLife.42496.

52. Hodcroft, E.B., Dyrdak, R., Andrés, C., Egli, A., Reist, J., García Martínez de Artola, D., Alcoba-Flórez, J., Niesters, H.G.M., Antón, A., Poelman, R., et al. (2022). Evolution, geographic spreading, and demographic distribution of Enterovirus D68. PLoS Pathog. *18*, e1010515. https://doi.org/10.1371/journal.ppat.1010515.

53. Pickett, B.E., Greer, D.S., Zhang, Y., Stewart, L., Zhou, L., Sun, G., Gu, Z., Kumar, S., Zaremba, S., Larsen, C.N., et al. (2012). Virus pathogen database and analysis resource (ViPR): A comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses *4*, 3209–3226. https://doi.org/10.3390/v4113209.

54. Olson, R.D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J.J., Dempsey, D.M., Dickerman, A., Dietrich, E.M., Kenyon, R.W., et al. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res. *51*, D678–D689. https://doi.org/10.1093/nar/gkac1003.

55. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2012). GenBank. Nucleic Acids Res. *41*, D36–D42. https://doi.org/10.1093/nar/gks1195.

56. Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schäffer, A.A., and Brister, J.R. (2017). Virus Variation Resource–improved response to emergent viral outbreaks. Nucleic Acids Res. *45*, D482–D490.

57. Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.

58. Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

59. Feng, Z., Xu, L., and Xie, Z. (2022). Receptors for respiratory syncytial virus infection and host factors regulating the life cycle of respiratory syncytial virus. Front. Cell. Infect. Microbiol. *12*, 858629. https://doi.org/10.3389/fcimb.2022.858629.

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Viral sequences	GISAID, ViPR-BRC, Genbank	Full list of sequence accession numbers and labs that contributed these sequences: https://github.com/blab/adaptive-evolution/blob/master/adaptive-evolution-analysis/manuscript_figures/acknowledgement_tables/sequence_acknowledgments.tsv
Software and algorithms		
Augur, version 14.1.0	Huddleston et al.,[44] https://doi.org/10.21105/joss.02906	https://github.com/nextstrain/augur
TreeTime, version 0.8.6	Sagulenko et al.[45] https://doi.org/10.1093/ve/vex042	https://github.com/neherlab/treetime
IQ-TREE, version 2.2.0	Nguyen et al.[46] https://doi.org/10.1093/molbev/msu300	https://github.com/Cibiv/IQ-TREE
Nextclade	Aksamentov et al.[47] https://doi.org/10.21105/joss.03773	https://clades.nextstrain.org
Nextstrain CLI, version 7.0.0	Hadfield et al.[48] https://doi.org/10.1093/bioinformatics/bty407	https://github.com/nextstrain
Custom code	This paper	https://github.com/blab/adaptive-evolution

RESOURCE AVAILABILITY

Lead contact

Inquiries for further information should be directed to the lead contact, Kathryn Kistler (kkistler@fredhutch.org).

Materials availability

This study did not generate any new reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The supplemental table provides accession numbers and contributing labs for each sequence used in this study.
- The code to implement the McDonald-Kreitman-based calculations of adaptation rates is located at https://github.com/blab/adaptive-evolution. All analysis code is written in Python 3 (Python Programming Language, RRID:SRC 008394) in Jupyter notebooks (Jupyter-console, RRID:SRC 018414). The results presented in this manuscript are also accessible in an interactive format at https://blab.github.io/ atlas-of-viral-adaptation/. Code to calculate rates of adaptation, as done in this study has been deposited on Mendeley Data (Mendeley Data: https://doi.org/10.17632/cj4n8m9kk4.1).
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

Input data for the estimation of rate of adaptation

The analyses in this manuscript require an alignment file in FASTA format, a tabular metadata file that contains the sampling date for each sequence in the alignment, and a reference file in Genbank format that supplies gene locations. We have written the analysis code with the intention that it should be paired with a Nextstrain build[48] (Nextstrain, RRID:SRC _018223), which will create the necessary alignment and metadata files as well as a phylogenetic tree, which is not necessary for the analysis but is a useful companion for the interpretation of the results. Of the pathogens considered in this manuscript, we used builds produced and maintained by the Nextstrain team for influenza A and B (https://github.com/nextstrain/seasonal-flu, created as part of[49]), measles (https://github.com/nextstrain/measles), mumps (https://github.com/nextstrain/mumps, created as part of[50]), dengue (https://github.com/nextstrain/dengue, created as part of[51]), and enterovirus D68 (https://github.com/nextstrain/enterovirus_d68, created as part of[52]). For the influenza A and B viruses, we tweaked the builds to contain sequences from any date, rather than limiting the tree to isolates sampled within the past 12 years. For dengue, we adjusted the pipeline to produce genotype-level phylogenies, rather than serotype-level. For all other viruses, we constructed a new Nextstrain build, as described below.

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

Sequence data

Norovirus, RSV, and rotavirus sequences were downloaded from ViPR/BV-BRC.[53,54] Adenovirus, hepatitis A, hepatitis B, parvovirus B19 genomes were downloaded from Genbank[55] Parainfluenza and seasonal coronavirus sequences were downloaded from both Genbank ViPR/BV-BRC and combined. Influenza C sequences were downloaded from NCBI Viruses.[56] All sequence queries were limited to clinical isolates from human hosts. All sequence data for a pathogen was curated into a single FASTA file, excluding sequences that had no available date information. Compiled sequence data for all these viruses are available via https://github.com/blab/adaptive-evolution.

Nextstrain builds to generate alignments and trees

Time-resolved phylogenies were generated for each pathogen by running a Nextstrain build.[48] Sequences were aligned to a reference genome using MAFFT[57] (RRID:SCR 011811). Trees were constructed using IQ-TREE[46] (RRID:SCR 017254), and branch lengths were inferred with TreeTime.[45] Builds were streamlined into a pipeline using Snakemake[58] (RRID:SCR 003475), and Snakemake workflows are available for each virus via https://github.com/blab/adaptive-evolution.

SARS-CoV-2 Nextstrain phylogenies

SARS-CoV-2 alignments and trees were retrieved from nextstrain.org builds. The build spanning all SARS-CoV-2 lineages contains samples from the beginning of the pandemic until February 13, 2023 and contains sequences evenly sampled over time and geography.

This dataset is viewable at https://nextstrain.org/ncov/gisaid/global/all-time/ 2023-02-26. The 21L-only build contains only sequences from the Omicron clade 21L up until April 9, 2023. This dataset is viewable at https://nextstrain.org/ncov/gisaid/ 21L/global/all-time/2023-04-18.

Rate of adaptation, with a fixed outgroup

The rate of adaptation within each gene of a genome is calculated from alignment of viral sequences sampled over time as in Bhatt et al.[30] To do this, the sequence alignment is broken up into constituent genes or subunits. Then, the gene-specific alignment is partitioned into 5-year windows tiling the entire span of time over which data is available. Windows are offset by 1 year so, for example, an alignment containing sequences from 1990-2022 would be partitioned into windows of [1990-1995, 1991-1996,..., 2016-2021, 2017-2022]. The exceptions are H1N1pdm and mumps where we use 3-year windows rather than 5-year, because there are only 12 and 17 years of data, respectively. The window size is a trade-off between picking up more signal (shorter windows), and reducing noise (longer windows) that can be due single sequences having a outsized effects on small sample sizes or chance sampling of one co-circulating clade over another. We require that each temporal window contain at least 3 isolates, and exclude time windows with 2 or fewer samples.

The outgroup sequence is found by taking a consensus of the sequences present in the first window. The choice to use a consensus sequence, rather than Most Recent Common Ancestor (MRCA), as the outgroup was based on previous implementations of this method,[28,30] to keep the method alignment- rather than phylogeny-based, and because, in our initial testing, similar rate estimates were obtained using MRCA or consensus outgroup. Each subsequent temporal window is then compared to the outgroup sequence to find polymorphisms and fixations. To do this, each nucleotide position in the gene alignment is compared to the outgroup to determine polymorphism, fixation, replacement and silent scores (see Bhatt et al.[29] and Bhatt et al.[30] for more details). The expectation for neutral evolution is found from the number of polymorphisms present at 15-75% and the number of silent (synonymous) fixations and near-fixations (greater than 75% frequency). The number of adaptive mutations within each window is calculated as the excess number of replacement (nonsynonymous) fixations or near-fixations above the neutral expectation. Ambiguously sequenced positions (N's) are ignored. The number of adaptive mutations are normalized by the gene length, and rates of adaptation are calculated as the slope of linear regression fitting adaptive mutations per codon over time. Bootstrap 95% confidence intervals were found by running the same method on 100 bootstrapped datasets. The bootstrapped datasets were created by sampling the codons in the outgroup, with replacement, and then applying the same codon order to the alignment.

Practically, the rate of adaptation is calculated using the rate of adaptation bhatt.ipynb notebook inside the adaptive-evolution-analysis/ directory, which reads in a virus-specific configuration file (in config/) that specifies necessary information to complete the analysis as well as metadata about the virus. For instance, the config files specify the relative locations of the necessary input data files, as well as which genes encode the polymerase and receptor-binding protein, whether the virus is enveloped, and what its primary mode of transmission is.

Rate of adaptation, with an updated outgroup

To account for viruses with especially high rates of evolution where multiple fixations have occurred at the same nucleotide position over the period of time the virus has been sampled, we update the outgroup sequence that is used for computing the rate of adaptation. The starting outgroup sequence is determined as with the 'fixed outgroup' method (explained above): as the consensus sequence of all isolates present in the first time window. Then, the outgroup sequence is updated each time a fixation (synonymous or nonsynonymous) occurs. Thus, future time windows are compared to an outgroup sequence that contains information about fixations that occurred in prior time windows. Simply overwriting the outgroup sequence at each fixation event allows more accurate determination of whether future mutations to the same nucleotide site or codon are synonymous or nonsynonymous. However,

Please cite this article in press as: Kistler and Bedford, An atlas of continuous adaptive evolution in endemic human viruses, Cell Host & Microbe (2023), https://doi.org/10.1016/j.chom.2023.09.012

because this site-counting McDonald-Kreitman based method estimates adaptive mutations in each time window by comparing the alignment to the outgroup, it is essentially counting the accumulation of all mutations that occurred between the outgroup and the time window, with no 'knowledge' of whether or not another mutation has previously occurred at any position. This means the method will only ever count a maximum of 1 fixation per nucleotide site. To make the counting method 'aware' of fixations that have occurred during previous time windows, the outgroup sequence is stored as a list, with the original outgroup sequence being the first element of the list and fixations getting added as subsequent list elements. At future timepoints, the method is thus 'aware' that a fixation has already occurred at any position where the outgroup sequence list has more than one element. The code to implement this method is in the notebook adaptive-evolution-analysis/rate of adaptation.ipynb.

Estimation of threshold, using logistic regression

To estimate the threshold of antigenic evolution, we ran a logistic regression predicting whether or not a virus is evolving antigenically (predictor variable) as a function of the estimated rate of adaptation in the receptor-binding protein (covariate). We used the receptor-binding proteins of three viruses that are known to evolve antigenically (influenza A/H3N2 HA1, influenza B/Vic HA1, and coronavirus 229E S1) and three that are known not to evolve antigenically (measles H, hepatitis A-IA VP1, and influenza C/Yamagata HEF1) in order to have an equal number of viral proteins in both categories for the logistic regression estimation. The threshold rate of antigenic evolution was then obtained as the rate at which the model assigns a greater than 50% probability of antigenic evolution (50% threshold for logistic regression analysis).

Rate of amino acid substitution

The rate of amino acid substitution was calculated from a phylogeny in order to account for repeated substitutions at the same position. For each virus, we traversed the phylogeny from root to tip, tallying the number of amino acid substitutions that occurred in the receptor-binding protein. Each tip was then plotted according to this accumulated number of substitutions and its sampling date. Linear regression of substitution count and time was used to calculate a rate of amino acid substitution for each virus.

Genes analyzed for each virus

Table S1 shows all genes used in the genome-wide analysis presented in Figure 3 and indicates their classification as receptor-binding, polymerase, surface-located (but not receptor-binding), and non-surface-located (but not polymerase). For some viruses, multiple proteins have been reported to have receptor-binding capacity in different strains or circumstances. For instance, influenza NA[36,37] and RSV F[59] proteins have been shown to bind receptors in some contexts. In these cases, we analyzed the canonical or primary *in vivo* receptor-binding protein. Viruses are listed in the order they appear in Figure 4.

QUANTIFICATION AND STATISTICAL ANALYSIS

In figures, circles show rates of adaptation estimated from empirical data and error bars show the 95% bootstrap percentiles, generated by estimating the rate of adaptation from 100 bootstrapped alignments. The threshold rate we use to infer whether a gene is undergoing ongoing adaptive evolution was determined by logistic regression using the rate of adaptation in the receptor-binding gene of 3 viruses known to evolve antigenically, and 3 viruses known to be antigenically stable.

ADDITIONAL RESOURCES

Estimated rates of adaptive evolution for each of the 28 viruses can be seen interactively at https://blab.github.io/atlas-of-viral-adaptation.