Virus Evolution, 2021, 7(2), 1–21
DOI: https://doi.org/10.1093/ve/veab073
Research Article

Decomposing the sources of SARS-CoV-2 fitness variation in the United States

Lenora Kepler,[1,†,‡] Marco Hamins-Puertolas,[2,†] and David A. Rasmussen[1,3,*,§]

[1]Bioinformatics Research Center, North Carolina State University, 1 Lampe Drive, Raleigh, NC 27607, USA, [2]Biomathematics Graduate Program, North Carolina State University, Campus Box 8213, Raleigh, NC 27695, USA and [3]Department of Entomology and Plant Pathology, North Carolina State University, Campus Box 7613, Raleigh, NC 27695, USA

[‡]http://orcid.org/0000-0002-7888-0517

[§]http://orcid.org/0000-0001-9457-7561

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: drasmus@ncsu.edu

Abstract

The fitness of a pathogen is a composite phenotype determined by many different factors influencing growth rates both within and between hosts. Determining what factors shape fitness at the host population-level is especially challenging because both intrinsic factors like pathogen genetics and extrinsic factors such as host behavior influence between-host transmission potential. This challenge has been highlighted by controversy surrounding the population-level fitness effects of mutations in the SARS-CoV-2 genome and their relative importance when compared against non-genetic factors shaping transmission dynamics. Building upon phylodynamic birth–death models, we develop a new framework to learn how hundreds of genetic and non-genetic factors have shaped the fitness of SARS-CoV-2. We estimate the fitness effects of all amino acid variants and several structural variants that have circulated in the United States between February 2020 and March 2021 from viral phylogenies. We also estimate how much fitness variation among pathogen lineages is attributable to genetic versus non-genetic factors such as spatial heterogeneity in transmission rates. Before September 2020, most fitness variation between lineages can be explained by background spatial heterogeneity in transmission rates across geographic regions. Starting in late 2020, genetic variation in fitness increased dramatically with the emergence of several new lineages including B.1.1.7, B.1.427, B.1.429 and B.1.526. Our analysis also indicates that genetic variants in less well-explored genomic regions outside of Spike may be contributing significantly to overall fitness variation in the viral population.

Key words: SARS-CoV-2; transmission fitness; phylodynamics; birth-death model

1. Introduction

Determining what factors shape the overall fitness of a novel pathogen such as SARS-CoV-2 is key to understanding the pathogen's epidemiological and evolutionary dynamics. However, quantifying pathogen fitness poses a number of conceptual as well as practical challenges. The fitness of a pathogen within a host, usually defined in terms of replication or growth rates, may only have a tenuous relationship with fitness at the host population-level, which is normally defined in terms of a pathogen's transmission potential (Handel and Rohani 2015; Xue and Bloom 2020). In addition to being scale-dependent, fitness is generally a composite phenotype determined by many different intrinsic (e.g. genetic) and extrinsic (e.g. environmental) factors. Several recent examples have highlighted how genetic mutations can dramatically increase the fitness of newly emerging viral pathogens including SARS-CoV, avian influenza and Ebola virus (Consortium et al., 2004; Long et al., 2016; Urbanowicz et al., 2016). At the same time, extrinsic factors such as climate and host behavior also strongly shape transmission dynamics and thereby pathogen fitness at the population-level (Shaman and Kohn 2009; Dalziel et al., 2018; Kissler et al., 2020). Studying fitness only on one scale, or only a single component of fitness, may therefore distort our overall picture of what factors most strongly determine pathogen fitness and transmission potential.

For SARS-CoV-2, reports of novel genetic variants with enhanced infectiousness or transmissibility emerged within the first months of the global pandemic and have since received considerable attention (Korber et al., 2020a; MacLean et al., 2020b; Tang et al., 2020). Early on, the most notable of these variants was the D614G mutation in the receptor binding domain of the Spike glycoprotein that binds human ACE2 receptors during cell entry. This variant spread rapidly around the globe in the spring of 2020 and apparently out-competed other viral genotypes that were already established in several locations (Korber et al., 2020b). Then in late 2020, several new variants of SARS-CoV-2 with increased transmissibility and potential antigenic escape mutations emerged, including lineage B.1.1.7 in the UK (Volz et al., 2021; Davies et al., 2021), B.1.351 in South Africa (Tegally et al., 2020) and P.1 in Brazil (Naveca et al., 2021). All of these variants were subsequently introduced into the United States as early as October or November, 2020 (Larsen and Worobey 2021; Washington et al., 2021). However, quantifying the fitness of these variants in the US and their impact on national-level epidemic dynamics poses a considerable challenge due to the rapidly evolving epidemic landscape in the US. In addition to introduced variants, new 'domestic' variants have emerged such as B.1.427/B.1.429 in California and B.1.526 in New York (Deng et al., 2021; Walensky, Walke and Fauci, 2021; Zhang et al., 2021). At the same time, older lineages like B.1.2 continued to dominant across large geographic regions even as new variants emerged (Pater et al., 2021). Furthermore, multiple lineages have independently acquired the same amino acid mutations suspected to increase transmission potential or escape immunity, including Spike E484K, Spike N501Y and Spike Q677P/H (Hodcroft et al., 2021; Martin et al., 2021), suggesting that lineages are adapting through convergent evolution.

While the fitness effect of genetic variants can be precisely quantified within hosts in controlled lab experiments (Urbanowicz et al., 2016; Muth et al., 2018; Zhang et al., 2020), laboratory conditions may not faithfully mimic within-host environments and immune responses encountered during natural infections. Moreover, due the scale-dependence of fitness, increased cellular infectivity or replication rates may not scale up to increase transmission potential between hosts, especially if within-host growth rates already produce sufficient viral loads or optimize a tradeoff between virulence and transmission (Fraser et al., 2007; Alizon et al., 2009; Ke et al., 2020). Thus, in order to provide a definitive answer about the epidemiological significance of a novel pathogen variant, we also need to quantify transmission fitness at the between-host level.

Transmission fitness at the between-host level can be inferred based on the evolutionary dynamics of pathogen variants in the host population. For example, the growth rate of alternate variants in a host population can be estimated from time series of variant frequencies or pathogen phylogenies as a surrogate for fitness (Foll, Shim, and Jensen, 2015; Kühnert et al., 2018). However, because fitness is a composite phenotype determined by multiple factors, inferring the fitness effect of a single feature such as a mutation can be easily confounded by other factors shaping pathogen fitness if these confounding factors are not accounted for. For example, a mutation of interest may be linked to other non-neutral mutations in the same genetic background and thereby confound estimates of the mutation's fitness effect by altering the background fitness of pathogen lineages carrying the mutation (Illingworth and Mustonen 2012; Neher 2013). Extrinsic factors such as climate and host behavior also strongly shape transmission dynamics (Dalziel et al., 2018; Kissler et al., 2020), such that a novel variant may increase rapidly in frequency and appear to have a fitness advantage simply by being in the right host population at the right time.

Viral phylogenies offer a promising way to estimate transmission fitness and disentangle the fitness effects of multiple genetic and extrinsic factors by tracking the genetic and non-genetic changes occurring along each lineage in the phylogeny. Here, we use the term lineage generally to refer to one or more branches in the phylogeny related by shared ancestry. On average then, a pathogen lineage with increased between-host fitness will be transmitted more frequently and have a higher probability of persisting through time. More fit lineages will therefore have a higher branching rate in the phylogeny and leave behind more sampled descendants. The fitness of a viral lineage can therefore be inferred from its branching pattern in a phylogeny using phylodynamic approaches such as birth–death models (Neher, Russell and Shraiman 2014). Multi-type birth-death (MTBD) models extend this basic idea by allowing the birth and death rate of lineages, and thereby fitness, to depend on a lineage's state or type, which may represent its genotype or any other *feature* representing a discrete character trait (Maddison, Midford and Otto, 2007; Stadler and Bonhoeffer 2013; Kühnert et al., 2018). Here we develop a phylodynamic inference framework that builds on earlier MTBD models to allow the fitness of a lineage to depend on multiple evolving traits or features (Rasmussen and Stadler 2019). In this framework, we first reconstruct ancestral states for all features that potentially predict fitness and then use a *fitness mapping function* to translate a lineage's reconstructed ancestral features into its expected fitness. We also develop a new approach that combines recent advances in machine learning with likelihood-based statistical inference under a birth–death model to learn this fitness mapping function from a phylogeny with reconstructed ancestral features.

We apply this new phylodynamic framework to learn what genetic as well as extrinsic features determined the transmission fitness of SARS-CoV-2 in the United States over the first year of the pandemic. This approach allows us to estimate the fitness effects of a large number of genetic variants while accounting for confounding factors such as background spatial heterogeneity in transmission. This approach also allows us to explore the relative importance of different features to overall transmission fitness by decomposing or partitioning fitness variation among lineages into parts attributable to different components of fitness. We therefore obtain a clearer picture of what factors have most strongly shaped the fitness of SARS-CoV-2 lineages circulating in the US.

2. Results

2.1 Phylogenetic and ancestral state reconstruction

We originally analyzed a data set containing 22,416 SARS-CoV-2 whole genome sequences sampled in the United States prior to September 1st, 2020 (pre-2020-09 data). Since the evolutionary dynamics of SARS-CoV-2 underwent a dramatic transition in late 2020, we subsequently performed an updated analysis on an additional 66,339 sequences sampled in the US between September 1st, 2020 and March 1st, 2021 (post-2020-09 data). We combine these two data sets (combined data) for some analyses below. Dated or time-calibrated maximum likelihood (ML) phylogenetic trees were reconstructed from whole genome sequences in each data set. For all sampled viruses, we also assembled a set of features that potentially predict fitness, including both genetic and non-genetic, environmental features. The genetic features include amino acid variants (AAVs) in coding regions spanning the SARS-CoV-2 genome as well as structural (deletion) variants. The non-genetic features include each sample's spatial location both at the level of US state and geographic region as determined by the US Department of Health and Human Services. Ancestral states for all features were then reconstructed for each node in the ML phylogeny. Thus, for each lineage in the phylogeny we obtain a vector of categorical variables representing ancestral features which we use to predict a lineage's fitness.

2.2 Background sampling and transmission heterogeneity

Because phylodynamic estimates will inevitably depend on what pathogens are sampled for genomic sequencing, we first estimated how sampling efforts varied across the US by time and geographic region. Sampling fractions were estimated based on

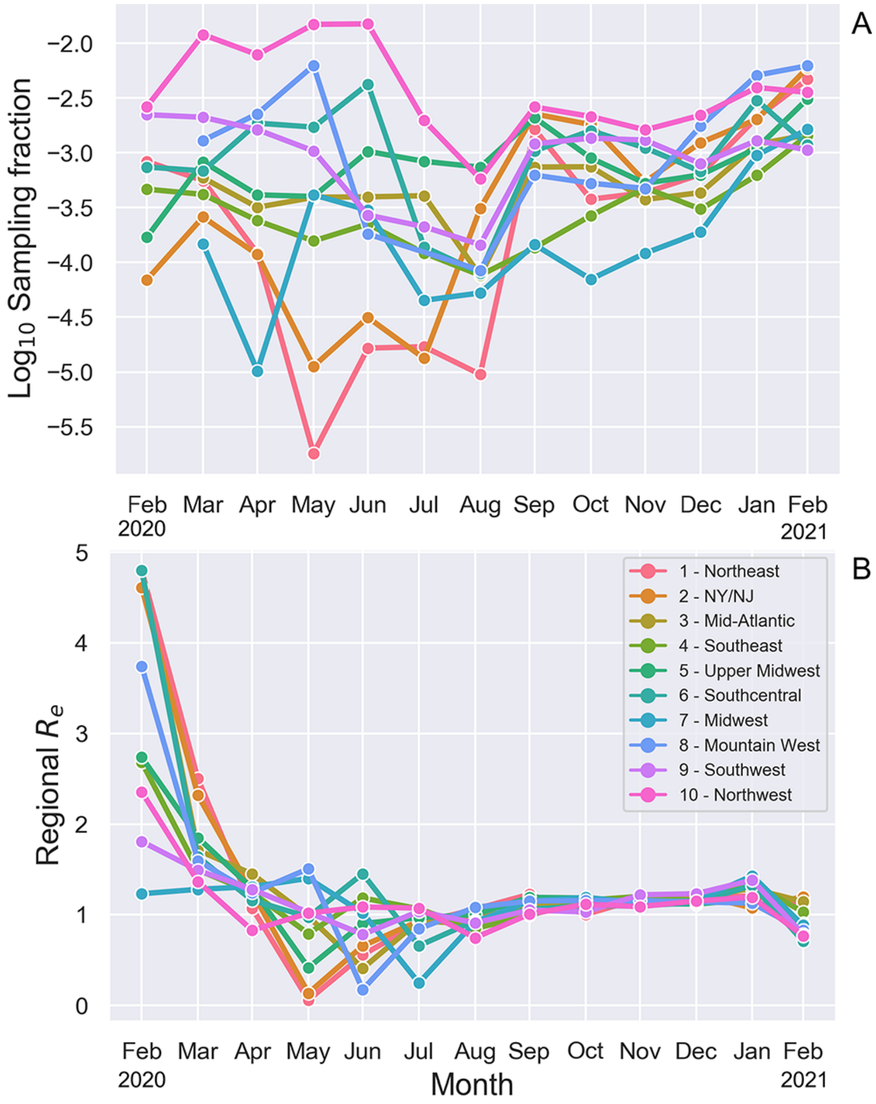

Figure 1. Background spatiotemporal heterogeneity in sampling fractions and effective reproductive number R_e of SARS-CoV-2 in the US. (A) Sampling fractions estimated based on the number of full viral genomes deposited to GISAID relative to the estimated number of total COVID infections in each region and time interval. (B) Effective reproductive number R_e estimates from the ML SARS-CoV-2 phylogeny. A regional transmission effect was estimated for each region and time interval, which was then used to rescale the estimated base transmission rate to compute R_e. The base transmission rate was estimated to be 0.184 per day, which assuming a constant recovery/removal rate of 0.14 per day yields an estimated time-averaged $R_e = 1.31$. States are grouped into the geographic regions designated by the US Department of Health and Human Services.

the number of whole-genome sequences submitted to GISAID relative to the total number of COVID infections imputed based on reported COVID deaths (see Methods). Overall, sampling fractions were extremely variable over the first eight months of the pandemic, but have become less variable and increased steadily over time since fall 2020 (Fig. 1A). When averaged across all times and regions, the mean sampling fraction is estimated to be 0.14%.

Before considering models that include genetic variants as fitness-predicting features, we considered several models accounting for background spatial and temporal variability in transmission, which could otherwise confound fitness estimates. The best fitting model allowed transmission rates to vary by both monthly time interval and geographic region (see Model Selection and Table 3). We therefore use a model that directly accounts for time-varying regional transmission rates and time-varying regional sampling fractions (as estimated in Fig. 1A) in all subsequent analysis. Note also that because the recovery rate and sampling fractions are treated as fixed parameters in subsequent analyses, all variation in fitness between lineages is attributable to variation in estimated transmission rates.

Using our best-fitting phylodynamic birth–death model, we estimated how background transmission rates varied across geographic regions from the SARS-CoV-2 phylogeny. Fig. 1B illustrates the changing transmission dynamics in terms of the effective reproductive number R_e for each region. Estimated transmission rates and R_e peak in February 2020, substantially earlier than peaks in reported cases. This pattern has been reported in other phylodynamic studies (Fauver et al., 2020; Nadeau et al., 2021; Ragonnet-Cronin et al., 2021), and may reflect considerable undetected transmission as well as lags in reporting before routine testing began. Transmission rates then remain low through the spring and early summer 2020 but are extremely variable across regions, likely reflecting the extreme variability in imputed sampling fractions during this same time period. Transmissions rates then steadily increase through late summer and fall of 2020 before declining in early 2021, consistent with trends observed in case report data.

2.3 Fitness effects of genetic variants

We next estimated the fitness effect of individual genetic variants while controlling for spatial heterogeneity in background transmission rates and sampling fractions. We consider the fitness effect of 66 AAVs in the pre-2020-09 data and 110 AAVs in the post-2020-09 data. However, in both data sets several variants are tightly linked and nearly always co-occur together

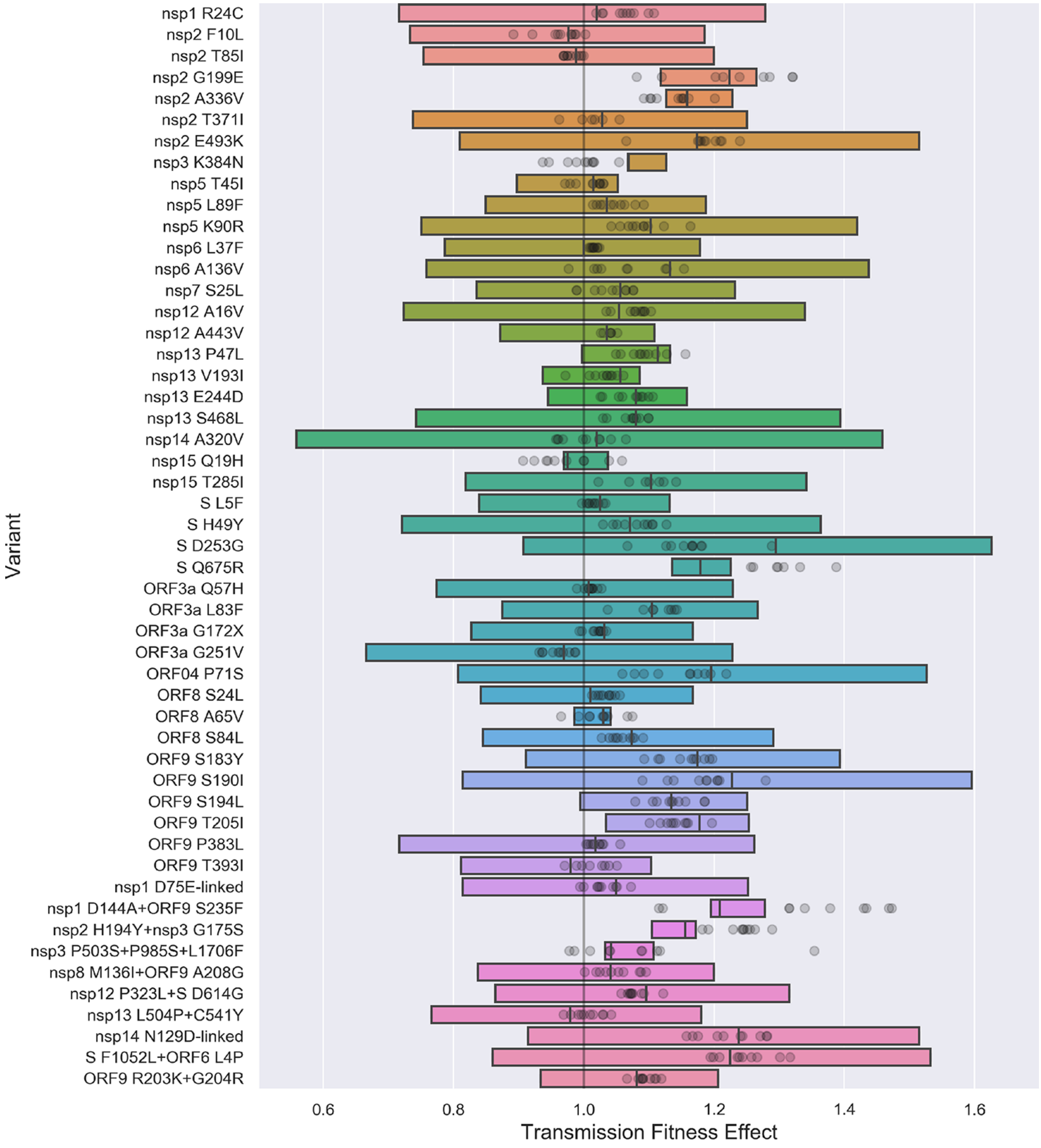

Figure 2. Estimated transmission fitness effects of amino acid variants in the pre-2020-09 data. Fitness effects are jointly estimated under a model of multiplicative fitness, such that neutral variants have a fitness of one. Variants are ordered from top to bottom by their genomic position. Vertical lines indicate the maximum likelihood estimate (MLE) and boxes reflect the extent of the 95% CI. The MLE of each fitness effect is also shown for ten replicate bootstrap trees as transparent circles. Sets of strongly linked variants are grouped together as single features to avoid collinearity among features. The nsp1 D75E-linked set includes nsp1 D75E, nsp3 P153L, nsp14 F233L and ORF8 V62L; the nsp14 N129D-linked set includes nsp14 N129D, nsp16 R216C, ORF3a G172V, ORF9 P199L and ORF9 P67S. The nsp14 N129D-linked set is referred to as the B.1.2 linked set below.

Table 1. Amino acid variants with significantly positive fitness effects in the pre-2020-09 data.

Variant	MLE	95% CI	Frequency
nsp2 A336V	1.15	1.13–1.22	0.006
nsp2 G199E	1.223	1.13–1.26	0.005
nsp3 K384N	1.068	1.06–1.12	0.015
nsp13 P47L	1.113	1.01–1.13	0.006
S Q675R	1.17	1.14–1.22	0.005
ORF9 S194L	1.13	1.01–1.24	0.048
ORF9 T205I	1.177	1.05–1.24	0.007

(Supplementary Fig. S1), leading to strong collinearity among features in our model. We therefore encode sets of linked variants with correlation coefficients greater than 0.95 as single features.

Fitness effects were estimated under a model where each variant has a multiplicative effect on the base transmission rate of a lineage such that a neutral variant has a fitness effect of 1.0 and deleterious or beneficial mutants have fitness effects less than or greater than 1.0, respectively. These fitness effects therefore also directly quantify the variant's effect on the R_e of lineages with the variant. We only consider a variant to be significantly deleterious or beneficial if the estimated 95% credible interval (CI) does not overlap with 1.0. A full list of all estimated fitness effects are available in Supplementary Data Files 1 and 2.

For the pre-2020-09 data, most AAVs are inferred to be neutral, with maximum likelihood estimates (MLE) of fitness effects close to 1.0 and 95% CIs overlapping 1.0 (Fig. 2). Variants with larger positive fitness effects (>1.05) are generally rare mutations or have wide confidence intervals surrounding the MLE. AAVs with large and significant positive effects are summarized in Table 1. Estimated fitness effects are generally consistent across 10 bootstrapped phylogeny replicates, indicating that our fitness estimates are not overly sensitive to the exact topology of the reconstructed ML tree.

The Spike D614G variant nearly always co-occurs with the P323L variant in nsp12 (RdRp), so we consider these two variants together as a single feature, but hereafter refer to this as the Spike D614G variant. Despite rapidly increasing in frequency in the spring of 2020 (Fig. 4A), the Spike D614G variant is estimated to have only a modest fitness benefit of 1.095 with a fairly wide 95% CI of 0.89–1.29. Simulations using a two-strain epidemiological model show that a transmission fitness effect of this magnitude is insufficient to explain D614G's rapid increase in frequency during the spring of 2020. Even if D614G entered the US through external introductions at a much higher rate than the ancestral 614D variant, D614G would have required a fitness advantage much larger than 10% to rise so rapidly in frequency (Supplementary Fig. S3). We therefore explore other plausible explanations for D614G's rapid rise below.

For the post-2020-09 data, most individual AAVs are again estimated to be approximately neutral (Fig. 3). Only one AAV, Spike A701V, is estimated to be significantly deleterious with a fitness effect of 0.937 (95% CI: 0.91–0.98). However, there are a relatively large number of AAVs with significant positive fitness effects between 1.05 and 1.10, especially in Spike as well as nsp3, ORF3a and ORF9 (Table 2).

Within the Spike domain, the putative antigenic escape mutation S E484K is estimated to have a significant fitness advantage (MLE: 1.117; 95% CI: 1.10–1.15). Two receptor binding domain mutations at position 501 with increased ACE2 binding avidity are also estimated to have significant positive fitness effects: S N501T (MLE: 1.091; 95% CI: 1.07–1.13) and S N501Y (MLE: 1.090; 95% CI: 1.05–1.11). The S Q677P variant, which has arisen in multiple genetic backgrounds (Pater et al., 2021; Hodcroft et al., 2021), linked with nsp6 Q160R is likewise estimated to have a significant fitness advantage (MLE: 1.109; 95% CI: 1.05–1.15), while the related mutation S Q677H is estimated to have a much smaller advantage (MLE: 1.026; 95% CI: 1.02–1.04). Finally, S P681H which has arisen multiple times including in B.1.1.7 and may aid cell entry by increasing the efficiency of furin cleavage (Garry et al., 2021), is estimated have a fitness effect of 1.069 (95% CI: 1.04–1.08).

To gain a better understanding of how the sampling of individual AAVs impacted estimated fitness effects, we explored how both the number of times each mutation was sampled and the number of times each mutation occurred independently in different lineages impacted inference. We found no apparent relationship between the number of samples and the MLE fitness effects (Pearson correlation coefficient $R = -0.009$, Supplementary Fig. S4A) and only a very weak negative correlation between the number of independent mutations and the MLE fitness effects ($R = -0.173$, Supplementary Fig. S4B), suggesting that the overall prevalence of each mutation did not unduly impact our fitness estimates. Likewise, there was only a weak negative correlation between the number of samples and the uncertainty surrounding our fitness estimates ($R = -0.116$, Supplementary Fig. S4C), although the mutations with the widest CIs were all rare mutations. Lastly, there was a slightly stronger negative correlation between the number of independent mutations and the width of the CIs ($R = -0.283$, Supplementary Fig. S4D), suggesting that our estimates became more confident for mutations that occurred independently many times in different genetic and spatial backgrounds.

Overall, the genetic features with the largest positive fitness effects in the post-2020-09 data are all sets of linked AAVs associated with major lineages. The B.1.526-linked variants are estimated to have the largest fitness effect (MLE: 1.322; 95% CI: 1.27–1.41), followed by the set of the nine B.1.1.7-linked variants (MLE: 1.183: 95% CI: 1.14–1.21) and then the B.1.427 and B.1.429-linked variants including Spike L452R (MLE: 1.159; 95% 1.08–1.229). Unfortunately, due to tight genetic linkage among these mutations, we are unable to determine whether individual AAVs within these linked sets contribute disproportionately to the fitness of these lineages.

2.4 Explaining the rapid rise of the Spike D614G variant

If the Spike D614G variant is not itself strongly beneficial as our fitness estimates suggest, what explains the rapid increase in the frequency of the D614G variant across the US? Stochastic processes including founder effects alone seem implausible given that the 614G variant appears to have out-competed and replaced the ancestral 614D variant even in geographic locations where the 614G variant arrived after the 614D variant (Korber et al., 2020b). We therefore consider two alternative hypotheses for the success of 614G: (1) the 614G variant gained an advantage by occurring in genetic backgrounds with higher fitness on average than the 614D variant; or (2) the 614G variant tended to occur in geographic locations with higher transmission rates on average.

We estimated the average background fitness of lineages with either the 614D or 614G variant, discounting the fitness effects of the Spike 614 variants themselves. Thus each lineage's background fitness reflects its geographic location as well as the fitness

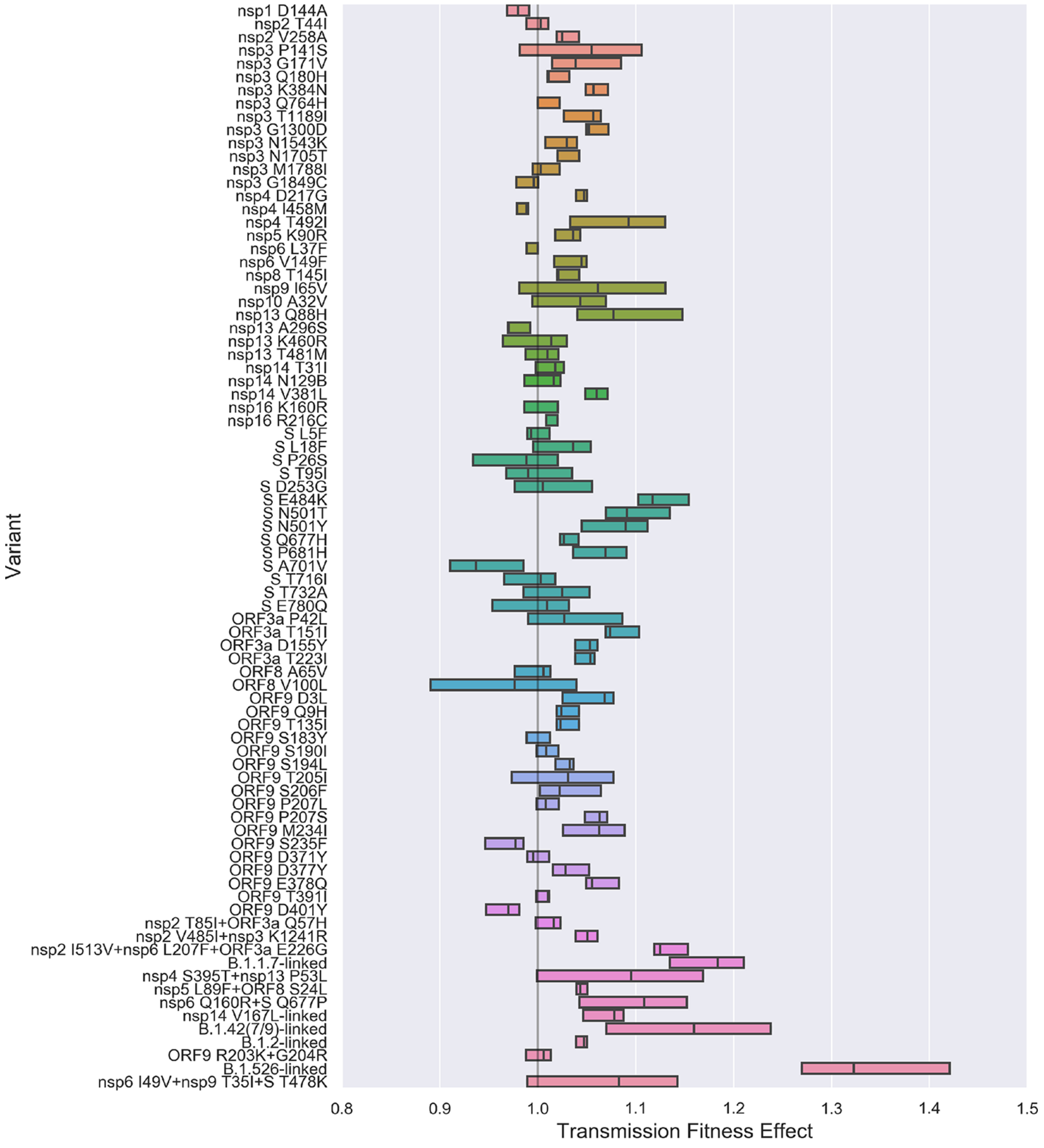

Figure 3. Estimated transmission fitness effects of amino acid variants in the post-2020-09 data. The fitness of each AAV is reported as a multiplicative effect on the base transmission rate. The B.1.1.7-linked set includes nsp3 T183I, nsp3 A890D, nsp3 I1412T, S A570D, S D1118H, S S982A, ORF8 Q27*, ORF8 R52I, ORF8 Y73C. The nsp14 V167L-linked set includes nsp12 V776L, nsp14 V167L, ORF3a S180P and ORF9 Q389L. The B.1.526-linked set includes nsp13 D260Y, S S13I, S W152C and S L452R. The B.1.2-linked set includes nsp14 N129D, ORF3a G172V, ORF9 P67S and ORF9 P199L.

effects of all genetic features besides the Spike 614 variants. Lineages with the 614G variant have an average background fitness that is 10.6% higher than the 614D variant. After partitioning total background fitness into genetic and spatial components, the 614G variant occurs in genetic backgrounds with 7.1% higher fitness. The genetic background fitness advantage of 614G lineages derives mostly from the ORF8 S84L variant, which we estimate had a fitness effect of 1.073 but with a high degree of uncertainty (95% CI: 0.87–1.26). However, the ORF8 S84L variant almost always occurs in the same genetic background as D614G, so it is unclear whether this fitness advantage should be attributed exclusively to S84L or to the overall genetic fitness background of lineages with 614G.

In addition to its genetic background, the 614G variant occurs in spatial backgrounds (i.e. geographic regions) with 3.4% higher transmission rates on average, although this average conceals the

Table 2. Amino acid variants with significantly positive fitness effects in the post-2020-09 data.

Variant	MLE	95% CI	Frequency
nsp3 K384N	1.057	1.05–1.07	0.016
nsp3 T1189I	1.057	1.03–1.06	0.012
nsp3 G1300D	1.052	1.05–1.07	0.023
nsp4 T429I	1.093	1.04–1.13	0.019
nsp13 Q88H	1.077	1.04–1.14	0.012
nsp14 V381L	1.06	1.05–1.07	0.014
S E484K	1.117	1.10–1.15	0.016
S N501T	1.091	1.07–1.13	0.011
S N501Y	1.090	1.05–1.11	0.026
S Q677H	1.026	1.02–1.04	0.054
S P681H	1.069	1.04–1.09	0.095
ORF3a T151I	1.074	1.07–1.11	0.015
ORF3a D155Y	1.053	1.04–1.06	0.018
ORF3a T223I	1.053	1.04–1.06	0.033
ORF3a E226G	1.124	1.12–1.15	0.014
ORF9 D3L	1.068	1.03–1.08	0.015
ORF9 P207S	1.063	1.05–1.07	0.015
ORF9 M234I	1.063	1.03–1.086	0.054
ORF9 E378Q	1.055	1.05–1.08	0.012

fact that the spatial background fitness advantage of the 614G variant was initially more than 10% due to first spreading in geographic regions with higher average transmission rates during the earliest stages of the pandemic, but this spatial advantage dissipated over time (Fig. 4). Directly comparing phylogenies with reconstructed ancestral states for the 614 variants with ancestral geographic locations makes clear that lineages carrying the 614G variant tended to be in locations like Region 2 (NY and NJ) and Region 5 (upper Midwest) with the highest transmission rates during the earliest stages of the pandemic (Supplementary Fig. S5).

The above analysis suggests that while lineages carrying the 614G variant may have had a small genetic fitness advantage, the 614G variant's rapid rise in frequency across the US was largely driven by establishing first in regions with higher average transmission rates. This can be seen by comparing the cumulative number of branching events in the phylogeny for lineages with the 614D or 614G variant. Using branching events as a proxy for transmission events, lineages with the 614G variant branch more often first in Region 2 and then subsequently in all other regions (Supplementary Fig. S6). Nevertheless, this pattern alone does not necessarily imply that the 614G variant has an intrinsic fitness advantage or elevated transmission rate as the 614G variant is also imported more frequently into each region than the D variant (Supplementary Fig. S7). To place the variants on more equitable footing, we therefore compare the branching/transmission rate of the variants *per lineage*, which accounts for the fact that the total number of lineages with the 614G variant in a given region may be higher due to either a higher transmission rate or importation rate. Contextualizing variant dynamics in this way, it becomes very clear that neither variant has a consistently higher branching rate through time (Fig. 5), supporting our model-based inference that the 614G variant may not have a major intrinsic fitness advantage. Averaging over all regions and time intervals up to May 1st, after which the 614D variant is rarely sampled, the branching rate of the 614G variant (mean = 0.13 per week) is 13% higher than the 614D variant (mean = 0.115 per week), consistent with our model-based estimate of a ∼10% fitness advantage, but these means are not significantly different (Welch's *t*-test = −1.42; *P*-value = 0.15).

2.5 Fitness of major lineages circulating in the US

Because the genetic features with the largest fitness effects are all linked sets of AAVs associated with individual lineages, we also estimated the fitness of major lineages circulating in the US as of March, 2021. Assignment of viruses to named lineages is based on the PANGO nomenclature system (Rambaut et al., 2020). For our purposes, we only consider PANGO lineages with at least 1,000 samples in the post-2020-09 data, plus B.1.526 ($n = 797$). Note, that this excludes some variants of concern that remained rare in the US, including the B.1.351 and P.1 variants.

Similar to the AAVs considered above, we estimated a multiplicative fitness effect on the base transmission rate for each lineage (Supplementary Fig. S8). Fig. 6 shows how the transmission fitness of these lineages fluctuated over time. The B.1 lineage carrying the Spike D614G variant has a clear fitness advantage over the ancestral A.1 lineage, but the fitness of both lineages declines rapidly during the spring of 2020. B.1.2 in turn has a moderate transmission fitness advantage over B.1, although the large peak in B.1.2's fitness relative to B.1 during the summer of 2020 is mostly attributable to a spatial transmission advantage of occurring in the Upper Midwest and other geographic regions with elevated transmission rates. The fitness of all major lineages then increases again in late 2020 just as four lineages with successively larger fitness advantages emerge. The rank order of these four lineages' fitness follows the same order as their associated sets of linked AAVs. B.1.427 and B.1.429, two sister lineages that first appeared in California and carry the Spike L452R mutation (Zhang et al., 2021), are estimated to have a transmission fitness effect of 1.232 (95% CI: 1.14–1.30) and 1.211 (95% CI: 1.13–1.28), respectively. B.1.1.7 has the second largest transmission fitness effect of 1.318 (95% CI: 1.27–1.35). B.1.526, which has spread rapidly in New York and carries the suspected antigenic escape mutation Spike E484K, is estimated to have the largest estimated transmission fitness effect (MLE: 1.397, 95% CI: 1.35–1.49) of all lineages.

The fitness advantage we estimate for B.1.1.7 is much smaller than the 50–70% increase in transmissibility estimated for B.1.1.7 in the UK (Davies et al., 2021; Volz et al., 2021). However, in the US, B.1.1.7 did not have the same explosive growth as it did in the UK and remained at relatively low frequencies in early 2021 despite arriving in the US as early as October or November, 2020 (Larsen and Worobey 2021; Washington et al., 2021). Nevertheless, to ensure our phylodynamic model is not underestimating the fitness of B.1.1.7, we also estimated the fitness of B.1.1.7 from a phylogeny of 30,000 SARS-CoV-2 genomes sampled in England between Sept 1st, 2020 and Feb. 1st, 2021. In England, we estimate the transmission fitness effect of B.1.1.7 to be 1.634 (95% CI: 1.61–1.65) relative to B.1, on par with earlier estimates (Supplementary Fig. S9). Thus, using the parental lineage B.1 as a basis for comparison, we estimate that the fitness of B.1.1.7 is 63% higher than B.1 in England but only 32% higher in the US.

Moreover, if anything, B.1.1.7 and other newly emerging variants are likely over-represented in the GISAID database due to preferential sequencing of variants of concern/interest. In particular, it is suspected that B.1.1.7 and other lineages with the Spike ΔH69/V70 deletion mutation were preferentially selected for sequencing because this deletion leads to Spike gene

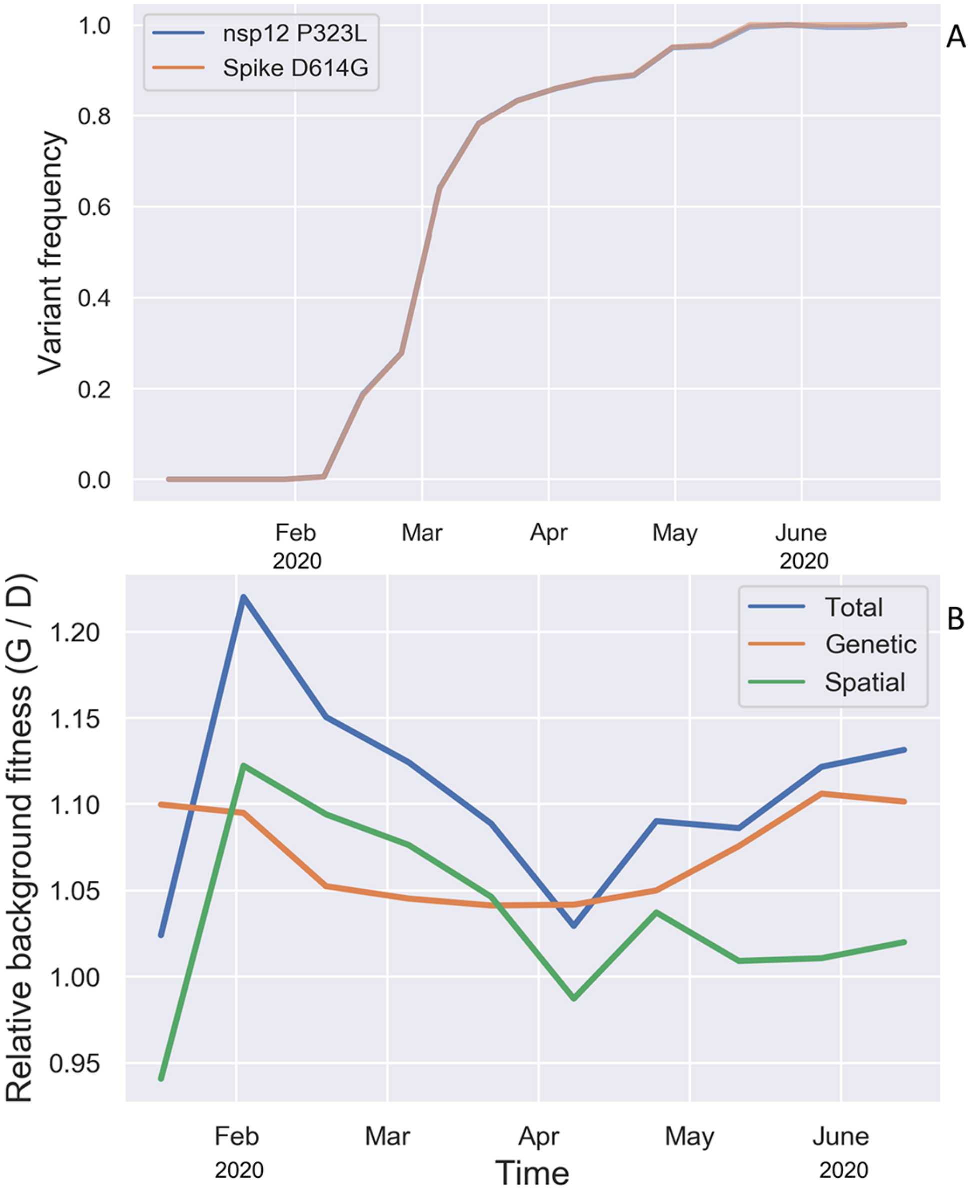

Figure 4. Evolutionary dynamics and background fitness of the Spike 614 variants. (A) Frequency of lineages carrying the Spike D614G and nsp12 P323L variants over time relative to all lineages in the ML phylogeny. These two variants are tightly linked so that they largely share the same evolutionary trajectory. (B) Relative background fitness of lineages with the Spike 614G variant versus the 614D variant. Background fitness was computed by averaging the fitness of all lineages with either variant present in the ML phylogeny at each time point. Total background fitness was then further split into a spatial and genetic component. Relative fitness is only shown up to July 1st, 2020 as the 614D variant was not sampled after this date.

target failure (SGTF) during diagnostic qPCR testing (Washington et al., 2020). Systematic oversampling of SGTF-associated variants with ΔH69/V70 could severely bias our phylodynamic fitness estimates. Indeed, a sensitivity analysis shows that the estimated fitness of B.1.1.7 declines exponentially as the assumed sampling fraction of B.1.1.7 increases (Supplementary Fig. S10).

Because other lineages besides B.1.1.7 share the ΔH69/V70 deletion and are also likely oversampled, we consider models that allow lineages with ΔH69/V70 to have their own SGTF-specific sampling fraction. SGTF-specific sampling fractions were estimated based on the number of GISAID sequences with the ΔH69/V70 deletion relative to the total number of COVID infections caused by a ΔH69/V70 variant. The later was imputed based on the number of SGTF-positive samples relative to all positive COVID tests using Helix's nation-wide diagnostic qPCR testing data (see Methods). Nationally, we estimate that SGTF samples were oversampled 4.11-fold, although there is extreme spatiotemporal heterogeneity in estimated sampling fractions (Supplementary Fig. S11). Regionally, we estimate that SGTF variants were oversampled by less than fourfold in most regions, but there were larger than tenfold sampling biases in the Southeast (Region 4) and Southwest (Region 9).

Accounting for SGTF-specific sampling fractions in our phylodynamic model by fixing the sampling fractions for the ΔH69/V70 variants at their estimated values did not substantially alter the estimated fitness of PANGO lineages (Supplementary Fig. S8). For lineages with the ΔH69/V70 deletion mutation, we estimate slightly lower transmission effects for B.1.427/B.1.429 but observe no change for B.1.1.7, suggesting that

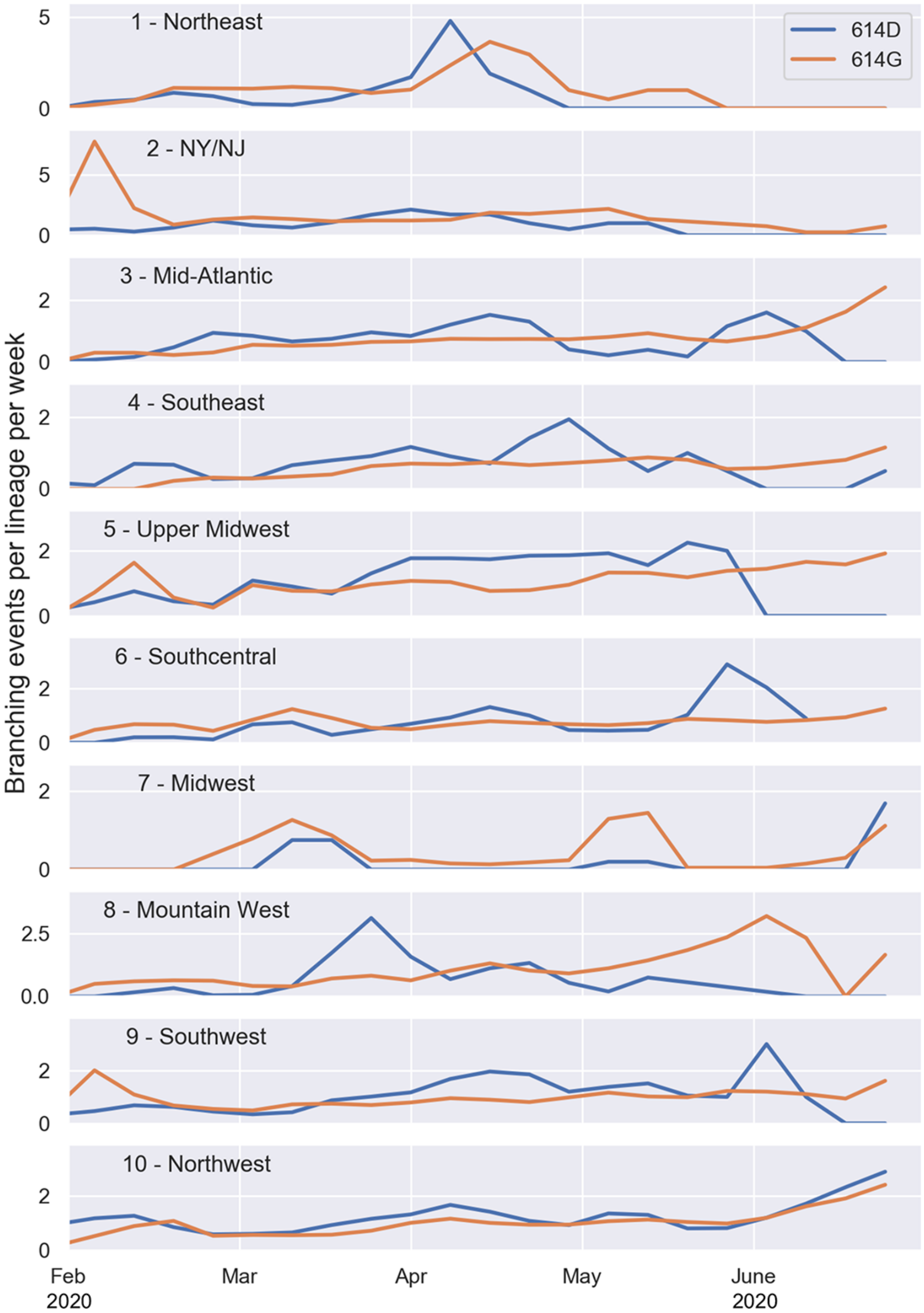

Figure 5. Branching rates per lineage in each region for lineages with the Spike 614D versus 614G variant. Branching rates are reported here as the number of branching events per week in the ML phylogeny for lineages with either variant.

our fitness estimates are largely robust to SGTF-specific sampling biases.

2.6 Decomposing the sources of fitness variation

Finally, we fit a model that included genetic features, spatiotemporal effects and branch-specific random effects to account for additional fitness variation not attributable to any feature or modeled source of variation in the model. Fitting this model to the SARS-CoV-2 phylogeny yields a fitness mapping function that we can use to predict the fitness of all lineages in terms of their transmission rate (Fig. 7).

Given the fitness of each lineage, we can compute how much fitness varies between lineages and then decompose total fitness variation into parts attributable to different components of fitness (see Methods: Decomposing fitness variation). At the beginning of the pandemic, virtually all fitness variation is attributable to spatial heterogeneity in transmission among geographic regions or to random effects which cannot be explained by spatial or genetic features in our model (Fig. 8B). As expected, genetic variants explain little to no fitness variation at the beginning of the pandemic when the virus population was genetically homogeneous. However, the fraction of fitness variation attributable to

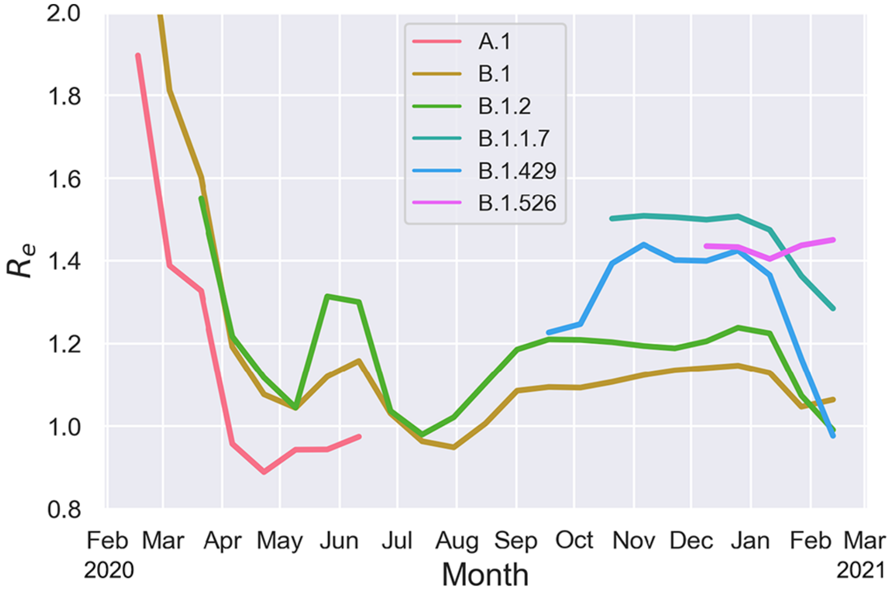

Figure 6. Time-varying transmission fitness of major PANGO lineages circulating in the US over time. The fitness of PANGO lineages is quantified as the average effective reproductive number R_e of all branches present in the phylogeny at each time point belonging to each PANGO lineage.

genetic variation quickly rises and then falls with the rise of Spike D614G in spring 2020. Genetic fitness variation then rises again in late summer and plateaus in late 2020, during which time approximately 30% of fitness variation is explained by genetic variants. Genetic variation in fitness then declines in early 2021 as less fit lineages are gradually replaced by more fit variants. Finally, a growing fraction of fitness variation through time is explained by random fitness effects, suggesting that an increasing fraction of fitness variation may be attributable to features not included in our model or that there may be more complex interactions among modeled features (e.g. epistatic interactions among mutations) that cannot be captured by our fitness mapping function.

Most genetic fitness variation is in turn attributable ta a small subset of genetic features including AAVs in Spike and sets of AAVs linked to major PANGO lineages (Fig. 8C). We also included AAVs in nsp3, ORF3a, and ORF9 as we found these coding regions contain several AAVs with large positive fitness effects (Fig. 3). Unexpectedly, AAVs in ORF3a and ORF9 at times contribute more fitness variation than AAVs in Spike, although this occurs mainly during the fall of 2020 when overall fitness variation is low. Nevertheless, this does suggest that less well-characterized regions of the genome outside of Spike may be shaping viral fitness in ways that remain poorly understood.

3. Discussion

Determining what factors shape viral fitness variation at the between-host level poses a major challenge to understanding a pathogen's transmission dynamics more generally. We therefore developed a new phylodynamic framework for learning how a large number of genetic and non-genetic features shape the overall fitness of SARS-CoV-2 at the host population-level. A major advantage of this framework is that it allows us to decompose or partition total fitness into different fitness components to learn how both intrinsic and extrinsic factors shape viral fitness. Applying this framework to over 88,000 viral whole genomes sampled in the United States over the first year of the pandemic, our results suggest that fitness variation among lineages was largely attributable to spatial heterogeneity in background transmission rates during the first months of the pandemic. The ability to partition fitness components between intrinsic genetic and non-genetic factors even revealed that the rapid rise of Spike D614G was due, at least in part, to a large spatial transmission advantage.

Before the emergence of several more fit variants in late 2020, we found that extrinsic, non-genetic factors like spatial heterogeneity in transmission rates consistently contributed more to overall variability in transmission fitness than viral genetic variation. Given that human mobility and non-pharmaceutical interventions such as social distancing appear to explain considerable variation in transmission rates within and between communities (Flaxman et al., 2020; Kissler et al., 2020; Chang et al., 2021), we strongly suspect that these same behavioral variables underlie the spatial transmission heterogeneity we infer from phylogenies. Unfortunately, the spatial resolution of our phylogenetic analysis was limited to geographic regions or at best US states. If we were able to track the movement of lineages with finer spatial resolution at the scale of individual communities where changes in human behavior appear to be most strongly correlated with reported cases, we could likely quantify how changes in human mobility or other behaviors shape transmission rates from phylogenies. We believe that this is an important direction for future work, as it would provide an independent means of measuring the impact of public health interventions on transmission rates using increasingly abundant pathogen sequence data (Ragonnet-Cronin et al., 2021; Rasigade et al., 2020).

However, since late 2020, an increasing fraction of fitness variation in the US is attributable to emerging variants, including the PANGO lineages B.1.1.7, B.1.427/B.1.429 and B.1.526. Although quantifying the relative fitness of these lineages is complicated by sampling biases, we estimate that these lineages have major transmission fitness advantages over earlier circulating lineages, and that our phylodynamic fitness estimates are largely robust to variant-specific sampling biases such as SGTF. In particular, we estimate that the 'domestic' variants B.1.427/B.1.429 and B.1.526 have a 20% and 40% fitness advantage relative to the ancestral

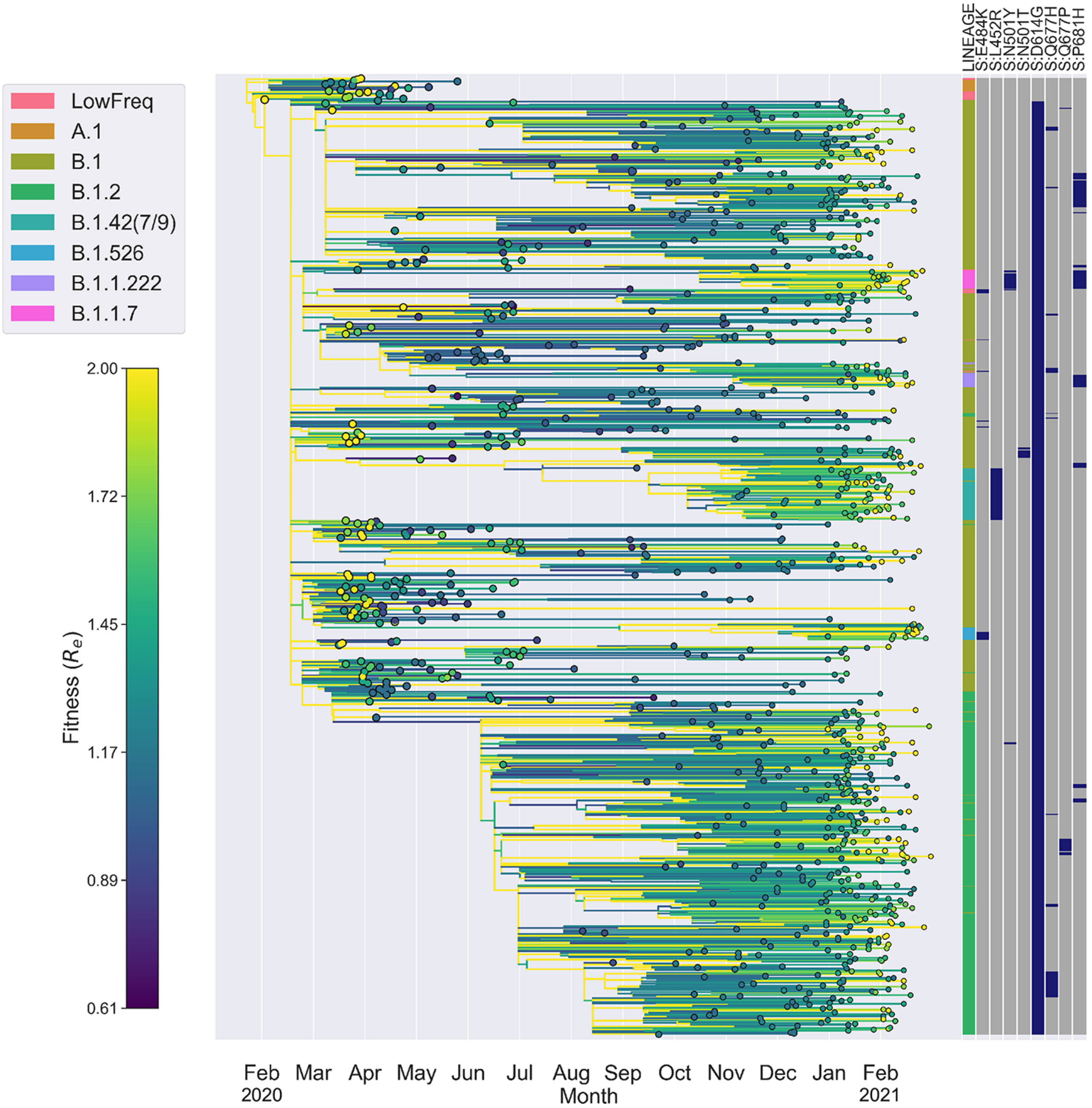

Figure 7. SARS-CoV-2 phylogeny with lineages colored by their transmission fitness (R_e). The annotation block provides each sampled tip's PANGO lineage assignment and genotype for several 'landmark' genetic features in Spike. Fitness is predicted based on each lineage's ancestral features using the fitted fitness mapping function with spatial, genetic and random effects. For the purposes of visualization, the full ML tree was thinned to include only 1000 randomly sampled tips and the fitness color scale was capped at a $R_e = 2$ to emphasize variation in fitness surrounding the mean rather than the full range of fitness values.

B.1 lineage, respectively. While the large fitness advantage we infer for B.1.526 may have been partially driven by increasing sampling in New York and the Northeast where this variant first emerged, this type of sampling bias is accounted for by the time and region-specific sampling fractions built into our model. We therefore think this variant may have had a true fitness advantage, possibly driven by the antigenic escape mutation Spike E484K. While our estimate of B.1.1.7's transmission advantage of 32% is not substantially lower than previous estimates of a 35%–45% advantage in the US (Washington et al., 2021), these estimates are much smaller than the 50%–70% transmission advantage estimated for B.1.1.7 in the UK using epidemiological data (Davies et al., 2021), coalescent-based methods (Volz et al., 2021), and our own phylodynamic birth–death methods. This smaller fitness advantage of B.1.1.7 is however consistent with its less explosive growth in the US. Despite arriving in the US as early as October or November, 2020 (Larsen and Worobey 2021; Washington et al., 2021), B.1.1.7 remained at low frequencies in many regions until early spring 2021. B.1.1.7 may have also faced increased competition from nearly equally fit 'domestic' variants, such that its

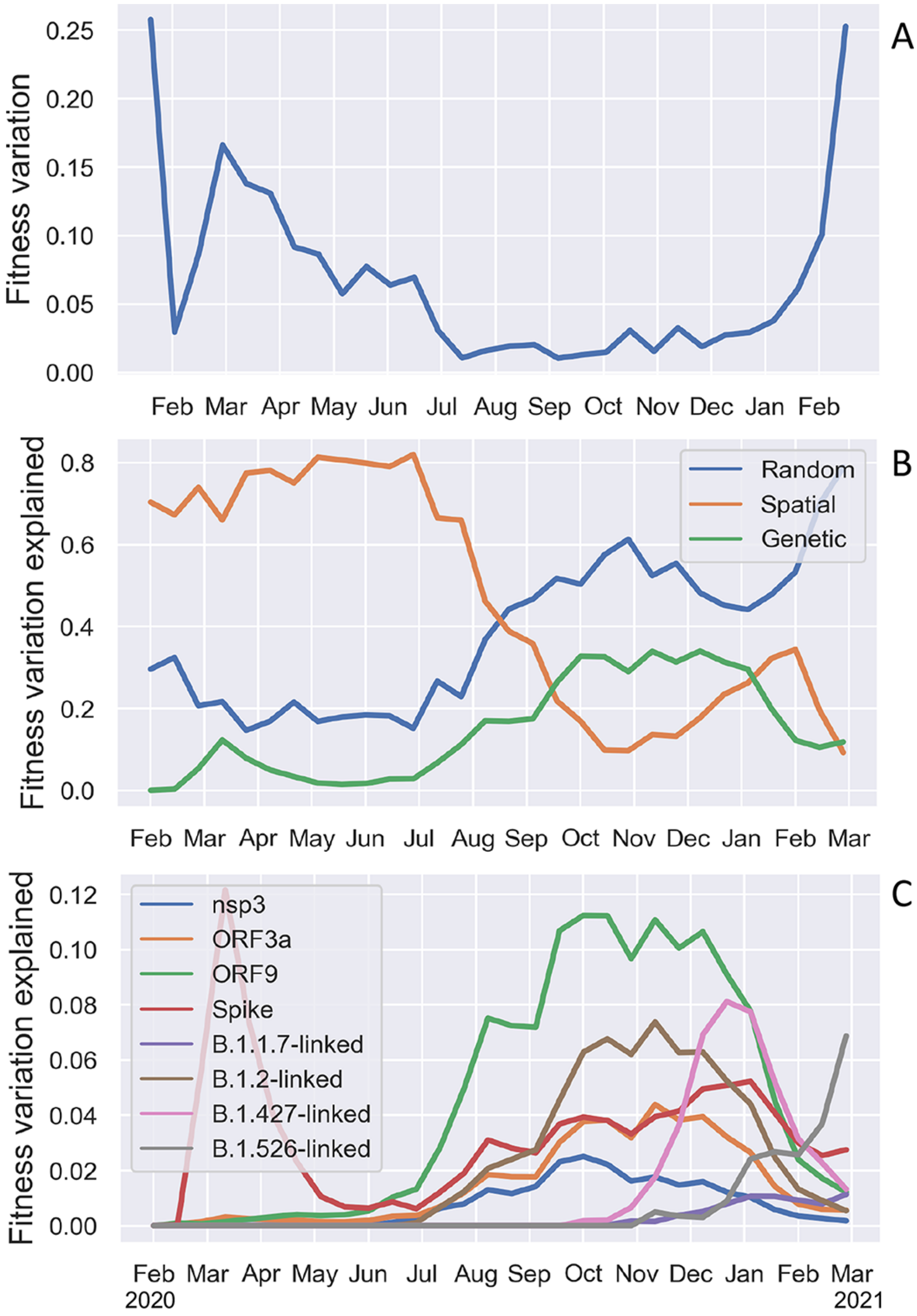

Figure 8. Fitness variation among lineages decomposed into sources attributable to different components of fitness. (A) Overall variation in fitness among lineages in the SARS-CoV-2 phylogeny through time. (B) Fraction of fitness variation explained by genetic, spatial and random fitness effects. (C) Fraction of fitness variation explained by different sets of genetic features.

growth rate in the US was impeded in an increasingly immunized population.

Another major advantage of our phylodynamic framework is that we can go beyond estimating the fitness of entire lineages and estimate the fitness effects of individual mutations while controlling for the fitness effects of other linked mutations. Exploring a large number of AAVs across the entire SARS-CoV-2 genome revealed moderate to large fitness effects in both expected and unexpected regions of the genome. As expected, we found several AAVs in Spike with substantial positive fitness effects on the order of 5%–10%. Most of these mutations are previously described variants that either increase cellular binding avidity (D614G), escape neutralizing antibodies (L452R and E484K) or both (N501Y/T) (Deng et al., 2021; Greaney et al., 2021; Zahradnik et al., 2021). Perhaps more surprisingly, we found several AAVs in coding regions outside of Spike with large positive fitness effects, including nsp3, ORF3a and ORF9. While these proteins remain less studied than Spike, they nonetheless play important roles in the viral life cycle or host-virus interactions. Nsp3 is the largest protein encoded by the SARS-CoV-2 genome and functions both as a protease and in anchoring the viral replication/transcription complex to cellular membranes (Lei, Kusov and Hilgenfeld, 2018). ORF3a is a multifunctional protein involved in cell membrane trafficking, host innate immune responses and apoptosis (Issa et al., 2020). Among bat and other non-human betacoronaviruses, ORF3a was found to have the greatest number of sites under (positive) episodic diversifying selection outside of Spike and the nucleocapsid, suggesting that it may facilitate host adaptation across species (MacLean et al., 2020a). ORF9 encodes the nucleocapsid (N), a key structural protein that is also immunogenic, although it is unclear if antibodies that recognize epitopes in N provide any neutralization potential (Gao et al., 2015; Ladner et al., 2021).

The general picture that emerges from our estimated fitness effects is that mutations across the entire SARS-CoV-2 genome are generally close to neutral but skewed towards small to moderate beneficial fitness effects at the between-host level. Previous work characterizing the fitness effects of de novo mutations suggests that a large fraction of mutations are expected to be neutral, but also suggests that most non-neutral mutations are deleterious rather than beneficial (Sanjuán, Moya and Elena, 2004; Eyre-Walker and Keightley 2007). This discrepancy between the fitness effects of de novo mutations and those circulating at the host population level is likely due to an ascertainment bias against the inclusion of deleterious mutations in our analysis, as these mutations would likely not have reached a frequency above our inclusion criteria of 0.5%. Another general trend that emerges based on the limited number of viral mutations in which fitness effects have been estimated both at the cellular or within-host level and at the between-host level is that the sign of the mutational fitness effects (beneficial or deleterious) tend to agree but the magnitude of fitness effects tend to be much smaller at the between-host scale (Rasmussen and Stadler 2019). Here, for example, we estimate much smaller fitness effects for Spike L452R, N501Y/T and D614G than their effects on cellular binding and infectivity would suggest (Starr et al., 2020; Deng et al., 2021; Zahradnik et al., 2021). We suspect that this trend towards mutations having more modest fitness effects at the between host level may arise simply due to the fact that an increasing number of processes influence fitness at higher scales. This could be analogous to how mutations in enzymes often have large effects on their kinetic activity but smaller effects on total metabolic flux if the enzyme impacted is not the rate limiting factor in a multi-step metabolic pathway (Kacser and Burns 1981). Similarly, mutations in Spike may increase binding avidity and thereby within host fitness, but may not have a proportional effect on between host fitness if cellular infection rates are not the process ultimately limiting transmission rates between hosts.

Spike D614G serves as an interesting case study to explore discrepancies in fitness estimated at the within and between scales. There is now considerable evidence that the Spike D614G mutation significantly alters viral fitness within individual hosts by increasing Spike's binding affinity to the human ACE2 receptor, more than doubling cellular infectivity and viral replication rates (Korber et al., 2020b; Plante et al., 2021). Higher viral replication rates could in turn explain why individuals infected with the 614G variant tend to have slightly higher viral loads (Wölfel et al., 2020; Korber et al., 2020b; Volz et al., 2021). Nevertheless, we estimate that D614G increases transmission fitness at the host-population level by only about 10%, which is at the low end of previous phylodynamic estimates from the UK which ranged from 10 to 29% (Volz et al., 2021). First, we note that increased replication rates and viral loads may not directly translate into increased infectiousness or transmission rates between hosts. While the relationship between viral load and infectiousness remains poorly understood for most respiratory viruses including SARS-CoV-2, recent work modeling clinical viral load data suggests that infectiousness does not increase linearly with viral load but with the logarithm of viral load such that it saturates at higher viral loads (Ke et al., 2020; Wölfel et al., 2020). This appears to fit a general pattern, as the amount of exhaled virus also saturates with increasing viral load for other seasonal coronaviruses (Leung et al., 2020). Given that the ancestral 614D variant was already able to efficiently replicate to high viral loads (Wölfel et al., 2020), it is conceivable that any additional replication advantage provided by the 614G variant would not significantly increase transmission rates further. The enhanced cellular infectivity and replication rates of D614G within hosts is therefore not irreconcilable with our inference that the mutant had more modest population-level fitness effect.

It does initially seem more challenging to reconcile the small estimated fitness advantage of D614G with its rapid spread and near universal rise in frequency around the world, especially as our simulations show that such a moderate fitness advantage would have been insufficient to explain the explosive growth of 614G observed in the US. Our phylogenetic analysis may partially explain this discrepancy, at least in the US. The ancestral 614D variant was largely limited to the West Coast (Regions 9 and 10), whereas the 614G variant established early in the Eastern US, especially in New York and New Jersey (Region 2). Due to the disproportionately large number of infections in New York and New Jersey during the early stages of the pandemic, the overall prevalence of the 614G variant also increased rapidly. Borrowing the concept of gene surfing from spatial population genetics may help to explain the rapid rise of the D614G variant. Gene surfing describes a scenario where a mutation can rapidly expand its geographic range by occurring along the edge or wave front of a spatially expanding population and then 'surf' to high frequencies by riding the wave of spatial expansion (Edmonds, Lillie, and Cavalli-Sforza, 2004; Klopfstein, Currat, and Excoffier, 2006; Hallatschek and Nelson 2008). While perhaps not a perfect metaphor here because SARS-CoV-2 did not spread as a spatially cohesive wave across the US, the gene surfing analogy captures the idea of how even a neutral mutation can be propelled to high frequencies across a range of spatial locations as a result of rapid population expansion. Viewed from this perspective, one can see why spatially aggregated time series of variant frequencies can be positively misleading about the fitness of a variant during a rapidly spatially expanding epidemic. Phylogenetic analysis coupled with ancestral state reconstructions offer a means of avoiding these pitfalls because they allow us to first identify lineages in the same transmission environment (e.g. geographic region) and then quantify the relative transmission rate of lineages from their branching pattern in the phylogeny.

Although some of the genetic features we considered were highly correlated with one another, relatively high mutation rates and viral movement rates between geographic regions means that most mutations were found in several different genetic and geographic backgrounds, providing us with sufficient information to disentangle the fitness effects of individual features. Several amino acid mutations also occurred repeatedly and independently in different lineages, which appears to have increased the precision of our fitness estimates (Supplementary Fig. S4D). How applicable our phylodynamic framework is to other pathogens beyond SARS-CoV-2 will therefore likely depend on the correlation structure among features being considered. If combinations of features are highly correlated or nearly collinear, such as if a mutation only occurs once in a single geographic context, it may not be possible to identify the fitness effects of individual features.

While our phylodynamic inference framework accounts for many potentially confounding factors including background variation in transmission rates, our analysis still has a number of limitations. First, inferences of pathogen fitness from phylogenies will inevitably depend on what lineages are sampled and included in the phylogeny. We can partially correct for sampling biases using time, location and even variant-specific sampling fractions but our fitness estimates will be dependent on the assumed sampling fractions. Second, our ancestral reconstructions did not account for external introductions or the possibility of ancestral lineages residing outside of the US. While neglecting introductions likely introduced minimal bias here due to the large size of the pandemic in the US, including global 'reference' samples collected outside of the US would have allowed us to identify and mask external lineages. Third, the computational efficiency of our approach relies on first reconstructing phylogenies and ancestral states before fitting our phylodynamic birth–death model. Although we did not do so here, we note that our fitness mapping functions do allow for ancestral features to be specified probabalistically to account for uncertainty in ancestral state reconstructions. Fourth, while we partially accounted for phylogenetic uncertainty by fitting models to replicate bootstrap phylogenies, using pseudoreplication to account for uncertainty is certainly a large step back from fully Bayesian phylodynamic methods that jointly infer key evolutionary and epidemiological parameters while simultaneously integrating over phylogenetic histories. Finally, we chose a simple fitness mapping function that assumes each feature has a multiplicative effect on lineage fitness. In reality, the relationship between a pathogen's genotype, environment and other features may be considerably more complex due to nonlinear relationships between features and fitness or interactions among genetic features (epistasis) and the environment (G×E interactions). It is therefore likely that some of the fitness variation attributed to random effects under our model are actually due to additional genetic sources such as epistatic interactions among mutations that cannot be captured under our simple multiplicative fitness model. Learning what types of functions are expressive enough to capture these complexities while remaining statistically tractable and biologically interpretable is a major challenge for future work.

As natural and vaccine-induced immunity continues to rise, new antigenic escape mutations and more transmissible variants like the B.1.617 (Delta) variant will likely continue to arise. Our phylodynamic framework can be used to examine the epidemiological significance of such mutations by estimating their transmission potential while accounting for confounding sources of fitness variation. Another major advantage of our approach is that it allows us to learn the relative importance of different features to overall pathogen fitness by decomposing fitness variation into its component parts. In the future, this will allow us to determine the contribution of new genetic variants relative to extrinsic factors such as host mobility. It may even be possible to tease apart fitness advantages driven by antigenic escape from increases in intrinsic transmission potential by incorporating antigenic fitness effects that depend on the immune profile of local host populations into the fitness mapping functions. Because fitness variation at the host population-level is essentially equivalent to variation in transmission potential, learning what features contribute the most to fitness variation is tantamount to learning what features most strongly regulate transmission. Thus, our phylodynamic learning framework not only allows us to estimate fitness, but understand what components of fitness shape both the evolutionary and epidemiological dynamics of viral pathogens.

4. Models and Methods

4.1 General approach

Our primary goal is to learn how multiple different character traits or *features*, which may include genetic variants, phenotypic traits and environmental variables, all act together to determine the fitness of pathogen lineages in a phylogenetic tree. We assume here that the phylogeny as well as ancestral features corresponding to the ancestral state of each feature is reconstructed beforehand. The relationship between predictive features and fitness is modeled using a *fitness mapping function* that predicts the expected fitness of a lineage based on its reconstructed ancestral features. The fitness mapping function can then be used to compute the expected fitness a lineage in terms of its birth and/or death rate. For a pathogen phylogeny, birth events are assumed to correspond to transmission events and deaths correspond to recovery or removal from the infected population. Given the birth and death rates of each lineage in a phylogenetic tree, the likelihood of the tree evolving as observed can be computed analytically under a birth–death-sampling model (Stadler 2009; Barido-Sottani, Vaughan, and Stadler, 2018). Our problem therefore reduces to finding the fitness mapping function that maximizes the likelihood of the phylogeny given the ancestral features of all lineages in the tree.

4.2 Phylogenetic reconstruction

For the original pre-2020-09 data set, a total of 22,416 SARS-CoV-2 whole genome sequences from the United States were downloaded from GISAID (Elbe and Buckland-Merrett 2017) on 2 October 2020 representing sequences that were sequenced prior to 1 September 2020. Genomes were aligned against a reference genome (NC_045512.2) using MAFFT version 7.475 (Katoh and Standley 2013). A ML phylogenetic tree was reconstructed in RAxML (Stamatakis 2014) using the rapid bootstrapping method with 10 bootstrap replicates assuming a GTR model of sequence evolution with Gamma-distributed rate variation among sites. The best ML and all bootstrapped trees were then dated using LSD (To et al., 2015) assuming a fixed clock rate of 0.0008 substitutions per site per year. A total of 93 sequences were discarded due to inconsistencies in sampling times or poor sequence quality.

Due to the large number of viral samples in the post-2020-09 data set, we extracted a ML phylogeny from a precomputed global SARS-CoV-2 ML phylogeny provided by Rob Lanfear's group on GISAID (Lanfear 2020). For all analyses conducted here, we use the 2021-03-13 version of the global tree. A focal tree containing all 66,339 viruses sampled in the US between 1 September 2020 and 1 March 2021 was then extracted using the *extract_tree_with_taxa* function in Dendropy version 4.5.1 (Sukumaran and Holder 2010). The extracted ML tree was then dated using LSD (To et al., 2015). For the fitness analysis of B.1.1.7 in the UK, we randomly sampled 30,000 viral isolates sampled in England between 1 September and 1 February. A focal tree containing these samples was then extracted from the global SARS-CoV-2 ML phylogeny.

4.3 Ancestral state reconstruction

In the original analysis of the pre-2020-09 data, ancestral states were reconstructed for each feature under a continuous-time Markov chain model of trait evolution using PastML (Ishikawa et al., 2019). PastML estimates the relative transition rate between each pair of states and the global (absolute) rate at which transitions occur. The relative transition rates are constrained to be proportional to the equilibrium frequencies of each state as under a F81 model of nucleotide substitution. Rate parameters were estimated independently for each feature. At each internal node, the state with the highest marginal posterior probability was taken to be the ancestral state for a given feature.

For the larger post-2020-09 data set, we reconstructed ancestral states using maximum parsimony (MP) to expedite analysis. This was motivated by the observation that there was typically little uncertainty surrounding the ML ancestral state reconstructions performed by PastML. For most genetic features, there was only a single or a small number of state transitions across the entire tree. In this case, MP and ML reconstructions are expected to agree as the most parsimonious and most likely reconstructions will be consistent given a slow rate of character evolution (Tuffley and Steel 1997; Zhang and Nei 1997). MP reconstructions were performed using Sankoff's dynamic programming algorithm (Sankoff 1975) implemented in Python.

Reconstructed ancestral features were then combined into a vector of categorical variables $\mathbf{x}_n$ for each lineage n. For categorical variables with more than one state, we used one-hot binary encoding to yield a strictly binary feature vector. Ancestral features were reconstructed for each bootstrap phylogeny independently.

4.4 Fitness mapping functions

Our main goal is to learn the fitness mapping function $F(\mathbf{x}_n)$ that maps the *features* of a lineage $\mathbf{x}_n$ to that lineage's expected fitness. Note that here, a lineage n will always refer to a single branch in the phylogeny. While $F(\mathbf{x}_n)$ could be any arbitrary function, we use a simple model that assumes the fitness effect β_i of each feature i is multiplicative:

$$F(\mathbf{x}_n) = \prod_{i \in \mathcal{X}} \beta_i \mathbf{x}_{n,i}, \tag{1}$$

where $\mathcal{X}$ is the set of all features used to predict fitness. Each feature $x_{n,i}$ is assumed to be encoded as a binary variable or as the probability of the lineage having a particular feature.

In order to decompose fitness into its component parts below we consider fitness effects on a log scale, which gives us the additive linear model:

$$\log(F(\mathbf{x}_n)) = \sum_{i \in \mathcal{X}} \log(\beta_i) x_{n,i}. \tag{2}$$

We also consider a fitness mapping function with random, branch-specific fitness effects u_n:

$$\log(F(\mathbf{x}_n)) = \sum_{i \in \mathcal{X}} \log(\beta_i) x_{n,i} + \log(u_n). \tag{3}$$

These random effects capture unmodeled sources of fitness variation such as genetic background effects at loci not included as features in the model.

Estimating branch-specific random fitness effects without additional constraints leads to extreme variability in fitness among lineages. In particular, long branches are estimated to have low fitness and short branches are estimated to have high fitness as this maximizes the likelihood of each branch under a birth–death model. We therefore use a Brownian motion model of trait evolution that constrains the branch-specific random fitness effects to be correlated between parent and child branches. Because each branch is assumed to have a unique random effect, we only allow fitness to change at branching events in the tree. The probability of a child having random fitness effect u_c given its parent's random fitness effect u_p is:

$$p(u_c | u_p, \Delta_t) = e^{-\frac{(u_c - u_p)^2}{2\alpha\Delta_t + \epsilon}}, \tag{4}$$

where Δ_t is the time elapsed between the parent and child node and α scales the variance of the child's fitness distribution. For numerical stability, we include a small value ϵ to ensure the probability does not become infinitely small when $\sigma\Delta_t << 1.0$. This model is conceptually similar to the ClaDS model of Maliet, Hartig and Morlon (2019) which estimates lineage-specific diversification rates by allowing for small shifts in birth and/or death rates at branching events, although the CLaDS model assumes a log-normal fitness distribution for child lineages independent of branch lengths.

How much fitness is allowed to vary between parent-child lineage pairs due to random effects is controlled by the hyperparameter α. We estimate α using k-fold cross-validation. Inspired by cross-validation techniques for time series data (Roberts et al., 2017), we longitudinally cross-section or block phylogenetic trees into training and test intervals. Random fitness effects are estimated for each branch in the tree during the training interval and then lineages in the test interval inherit their random fitness effects from their parent (or most recent ancestor) in the training interval. Thus, if the random fitness effects capture true fitness variation among lineages in the training interval, these fitness effects should more accurately predict the fitness of descendent lineages and improve the likelihood of the phylogeny in the test period. In contrast, a model with α set too high will overfit the fitness variation among lineages in the training period but will not improve performance in the test period. We can therefore use cross-validation to estimate an optimal value of α that maximizes the likelihood of trees in the test period while preventing the random fitness effects from overfitting fitness variation among lineages.

4.5 The phylodynamic birth–death-sampling model

The likelihood of a phylogenetic tree evolving as observed can be computed under a phylodynamic birth–death-sampling model (Stadler 2009) given the expected fitness of each lineage, which we predict based on a lineage's ancestral features $\mathbf{x}_n$ using a fitness mapping function $F(\mathbf{x}_n)$. We assume throughout that fitness is directly proportional to a lineage's birth or transmission rate $\lambda_n = f_n \lambda_0$, where λ_0 is a base transmission rate which is scaled by a lineage's fitness f_n. We also assume that the removal rate μ and sampling fraction σ are constant across all lineages, although we consider models where σ is allowed to vary by time and region below. This dramatically simplifies the model, as instead of having a multi-type birth–death process we have a series of connected single-type birth–death processes along lineages who's birth and death rates are piecewise constant.

Under this model, it is possible to analytically compute the likelihood of the phylogeny evolving as observed, allowing for efficient statistical inference. Given the birth, death and sampling rates and the fitness mapping function to compute the expected fitness of each lineage, the likelihood of each lineage or subtree evolving is independent conditional upon knowing the ancestral features used to predict fitness. The total likelihood of a phylogenetic tree $\mathcal{T}$ can be decomposed into the likelihood of a set of sampling events S, a set of branching (transmission) events B, and a set of lineages

N:

$$L(T|F(x),\lambda_0,\mu,\sigma) = \prod_{b\in B} L_{branch}(b) \prod_{s\in S} L_{sample}(s) \prod_{n\in N} L_{line}(n). \quad (5)$$

The likelihood of an individual branching or transmission event is:

$$L_{branch}(b) = F(x_{n(b)})\lambda_0 = \lambda_{n(b)}, \quad (6)$$

where we use the notation $n(b)$ to refer the parent lineage involved in a particular branching event b.

The likelihood of an individual sampling event at time t in the past is:

$$L_{sample}(s) = \begin{cases} \sigma\mu & \text{if } t > 0 \\ \rho & \text{if } t = 0. \end{cases} \quad (7)$$

Before the present, the probability of a sampling event depends on the removal rate μ and the probability σ that the lineage is sampled upon removal. At the present ($t=0$), any extent (i.e. currently infected) individual is sampled with probability ρ.

$L_{line}(n)$ gives the likelihood a lineage n evolved as observed; i.e. the probability that the lineage survived without giving rise to other sampled lineages. As shown in Barido-Sottani, Vaughan, and Stadler (2018), over a time interval of length Δ_t, this likelihood can be computed as:

$$D_n(\Delta_t) = e^{c\Delta_t}\left(\frac{y_n - x_n}{(y_n + \lambda_n E_n(t))e^{-c\Delta_t} - (x_n + \lambda_n E_n(t))}\right)^2, \quad (8)$$

with:

$$c_n = \sqrt{(\lambda_n + \mu)^2 - 4\mu(1-\sigma)\lambda_n}. \quad (9)$$

$$x_n = \frac{-(\lambda_n + \mu) - c}{2}, \quad (10)$$

$$y_n = \frac{-(\lambda_n + \mu) + c}{2}. \quad (11)$$

The $E_n(t)$ terms in (8) represent the probability that a lineage at time t in the past produced no sampled descendants. Assuming that the birth, death and sampling rates do not change along unsampled lineages from their values at time t, these probabilities are given by:

$$E_n(t) = -\frac{1}{\lambda_n}\frac{(y_n + \lambda_n E(0))x_n e^{-c_n t} - y_n(x_n + \lambda_n E(0))}{(y_n + \lambda_n E(0))e^{-c_n t} - (x_n + \lambda_n E(0))}. \quad (12)$$

$E(0)$ is the initial condition or probability of lineage not being sampled at the present ($t=0$). Given that the proportion of individuals sampled at present is ρ, we set $E(0) = 1 - \rho$. For simplicity we assume that $\rho = \sigma\mu/365$ so that the probability of a lineage being sampled on the final day of sampling is proportional to the probability of an individual being removed from the infectious population on that day, but the sampling fraction is the same as any point in the past.

4.6 Model fitting and statistical inference

Learning the fitness mapping function from a phylogenetic tree is a somewhat non-standard problem in that we do not have direct observations of a lineage's fitness to which we can compare our predictions under $F(x)$. Nevertheless, we can formulate statistical inference as an optimization problem where we seek to find the fitness mapping function $F(x)$ with parameters $\hat{\theta}$ that maximizes the overall likelihood of the phylogeny given the reconstructed ancestral features under the birth–death-sampling model:

$$\hat{\theta} = \arg\max_{\theta} L(T|F_\theta(x),\lambda_0,\mu,\sigma). \quad (13)$$

Formulating the problem in this way opens the way to using efficient optimization algorithms developed in recent years to train neural networks and other machine learning models. Instead of optimizing a typical loss function (e.g. least-squares), we simply maximize the likelihood of the phylogeny under the birth–death-sampling model. In particular, we use the ADAM optimizer (Kingma and Ba 2014), a form of stochastic gradient descent (SGD) which adapts learning rates based on gradients (i.e. first-order derivatives) of the likelihood function with respect to different parameters. Adapting the learning rates allows the algorithm to accelerate its momentum towards parameters that optimize the loss function. To make use of ADAM and other high-performance SGD algorithms, we implemented our fitness mapping function and birth–death likelihood function in TensorFlow 2 (Abadi et al., 2016). Gradients in the likelihood function are computed using TensorFlow's auto-differentiation functionality, allowing us to efficiently fit complex models with hundreds of features or parameters. Using this approach, even fitting our most complex model with over 300 free parameters to a phylogeny with over 22,000 tips only takes a few minutes on a standard desktop computer.

Learning the fitness mapping function through gradient descent provides MLEs of each parameter in the model. To quantify uncertainty surrounding the MLEs, we compute the likelihood of the phylogeny over a fixed grid of parameters values and then determine which values fall within the 95% CIs using an asymptotic chi-square approximation to the likelihood ratio test.

4.7 Performance on simulated data

To test the ability of our methods to correctly estimate fitness effects, we ran forward simulations where both genetic and spatiotemporal features influence viral fitness. Phylogenies were simulated under a birth–death-sampling model using the stochastic Gillespie algorithm (Gillespie 2007) starting with a single infected individual. In all simulations we assume a constant base birth rate of 1.2 and death rate of 1.0 per time unit. A virus's genotype is represented by ten binary sites where zeros indicate the ancestral state and ones indicate the mutant state. Each site has a random, multiplicative effect on fitness when mutated to the one state. Mutation occurs at a constant per site rate of 1.5×10^{-2} per time unit. A lineage's spatial location is encoded as an additional evolving character trait. To emulate background fitness variation due to spatiotemporal heterogeneity in transmission, each combination of region and time interval is assigned a background transmission rate. Individuals move from one region to another with a transition rate of 0.3 per time unit. Furthermore, in order to emulate an additional source of fitness variation not directly accounted for in the inference model, we added transmission heterogeneity by having each infected individual draw a random effect that rescales their transmission rate from a gamma distribution (Lloyd-Smith et al., 2005). Here the gamma distribution has a dispersion parameter 0.15 and scale parameter 10, such that on average, there is a branch-specific fitness of 1.5, but individually, fitness varied substantially.

Performance was tested under both a high sampling regime ($\sigma = \rho = 0.5$) and a low sampling regime ($\sigma = \rho = 0.05$). A phylogeny was built from the true ancestral history of sampled individuals. True ancestral features (states) were assumed to be known for the purposes of validating the inference algorithm. Simulations were run for 8 time units and those that ended more than 0.2 time units before then, or that had less than 800 sampled individuals, were discarded. We then estimated background transmission rates and genetic fitness effects from each simulated phylogeny.

Estimated spatiotemporal and genetic fitness effects are generally well correlated with the true values used in simulations (Supplementary Fig. S12). However, estimation accuracy depends largely on the overall sampling fraction and the number of individuals sampled with a given feature (spatial location or genotype). In particular, the fitness effects of rare features sampled at low frequencies tend to have the most variable and least accurate estimates. Because estimating the fitness of rare features under a birth–death model appears to be inherently difficult (Rasmussen and Stadler 2019), we only estimate fitness effects for features with a sampling frequency above 0.5% from empirical SARS-CoV-2 phylogenies.

We also tested our ability to estimate branch-specific random effects under the Brownian motion model (4). Phylogenies were simulated with site-specific fitness and random branch effects evolving with the variance scaling parameter α set to 0.05, resulting in considerable random fitness variation between lineages. Estimated random effects for each branch were tightly correlated with their true random effects (Supplementary Fig. S13A). Furthermore, accounting for branch-specific random effects improves estimates of site-specific fitness effects relative to a model without random branch effects (Supplementary Fig. S13B-C).

4.8 Birth–death-sampling model parameters for SARS-CoV-2

Because it is not possible to estimate all of the parameters in the birth–death-sampling model from a phylogeny alone (Stadler et al., 2013), we fix some parameters at values based on prior knowledge. We assume individuals infected with SARS-CoV-2 stay infected (and infectious) for 7 days on average, leading to a removal rate $\mu = \frac{1}{7}$ per day.

In several models we allow the base transmission rate λ_0 or sampling fraction σ to vary over time. In this case we have a time-varying transmission rate $\lambda(t)$ and $\sigma(t)$ that depends on the time t. However, this can easily be incorporated into the birth–death model above. If a lineage's transmission rate or sampling fraction changes along a branch due to an underlying change in $\lambda(t)$ or $\sigma(t)$, we simply divide the branch into segments corresponding to the time intervals over which these parameters remain piecewise constant and add each lineage segment to the set of lineages in N.

We assume that the sampling fraction σ was zero before the first sample in our data set was collected in January 2020. After the first sampling date, we allow the sampling fraction to vary by time and region as described below.

4.9 Modeling sampling heterogeneity

In order to estimate how sampling fractions vary across space and time, we count the number of sequence samples $g_{i,t}$ submitted to GISAID within each geographic location i over each time interval t. An unbiased estimate of the sampling fraction $\sigma_{i,t}$ would therefore be:

$$\sigma_{i,t} = \frac{g_{i,t}}{c_{i,t}}, \tag{14}$$

where $c_{i,t}$ is the total number of infections or cumulative incidence in region i over time interval t.

We of course do not know $c_{i,t}$ but can obtain a pseudo-empirical estimate $\hat{c}_{i,t}$ by considering the number of deaths attributed to SARS-CoV-2 $d_{i,t}$ and the estimated infection fatality ratio ϕ, which was assumed to be 0.5% (Perez-Saez et al., 2021). We can therefore approximate the total number of cases $c_{i,t}$ as:

$$\hat{c}_{i,t} = \frac{d_{i,t}}{\phi}. \tag{15}$$

Substituting $\hat{c}_{i,t}$ for $c_{i,t}$ in (14), we arrive at a crude estimate of the sampling fraction.

While the case fatality ratio likely also fluctuates over space and time due to changes in the age distribution of infections among other reasons, it seems reasonable to assume that the mortality rate fluctuates less than the testing or sequence sampling fraction (Flaxman et al., 2020). We can therefore roughly estimate the total number of cases based on the number of observed deaths. Using this approach, we estimate that there were a total of 35,134,400 cumulative cases in the US by September 1st, whereas the total number of positive cases reported by the COVID Project (https://covidtracking.com/data/national) on the same date was 6,017,826. Our estimate for the total number of cases suggests that 83% of all infections were not detected in the US, which is consistent with recent estimates by Wu et al., (2020), who estimated that up to 89% of all infections are unreported using an independent approach.

Using data from the COVID Project to tabulate cumulative deaths $d_{i,t}$ for each region and time interval, we estimate how the sampling fraction $\sigma_{i,t}$ varied across regions and time (Fig. 1A). For these estimates we assume reported deaths lag behind reported cases by three weeks when estimating sampling fractions.

4.10 Modeling variant-specific sampling biases

An additional complication arises here because B.1.1.7 and other lineages carrying the Spike ΔH69/V70 deletion mutation are likely oversampled due to preferential sequencing of viral isolates suspected of being newly emerging variants based on Spike gene target failure (SGTF) during diagnostic qPCR testing (Washington et al., 2020). To account for SGTF-related sampling bias, we estimate a SGTF-specific sampling fraction for lineages with the ΔH69/V70 deletion. Note that we estimate sampling fractions for all lineages with the ΔH69/V70 deletion rather than just B.1.1.7 as other lineages share this deletion and are therefore likely also preferentially selected for sequencing. We estimate SGTF-specific sampling fractions based on the fraction of SGTF-positive samples $\xi_{i,t}^{SGTF}$ relative to the total number of SARS-CoV-2 positive samples using Helix's nationwide diagnostic qPCR data (https://github.com/myhelix/helix-covid19db/blob/master/counts_by_state.csv). We then modify (14) by multiplying the total number of infections $c_{i,t}$ by the SGTF-positive fraction $\xi_{i,t}$:

Table 3. SGTF-specific sampling fractions (as percentages) and the ratio of SGTF versus non-SGTF sampling fractions (parentheses).

Region	Nov 2020	Dec 2020	Jan 2021	Feb 2021
Region 1	1.07 (24.8)	1.83 (28.8)	0.9 (4.31)	0.23 (0.50)
Region 2	0.1 (1.74)	0.42 (3.38)	0.61 (3.04)	0.62 (1.04)
Region 3	0.04 (1.07)	0.16 (3.68)	0.39 (3.24)	0.18 (1.26)
Region 4	0.08 (1.66)	0.50 (16.3)	0.62 (9.9)	0.12 (0.8)
Region 5	<0.01 (0.16)	0.15 (2.33)	0.37 (3.40)	0.31 (1.00)
Region 6	0.02 (0.15)	0.06 (0.85)	0.35 (1.18)	0.07 (0.57)
Region 7	–	0.03 (1.46)	0.18 (1.95)	0.15 (0.94)
Region 8	0.04 (0.87)	0.43 (2.45)	1.36 (2.67)	0.95 (1.53)
Region 9	0.45 (3.48)	0.93 (11.68)	1.46 (11.32)	0.32 (3.0)
Region 10	0.09 (0.55)	0.22 (1.0)	0.56 (1.41)	0.38 (1.06)

Table 4. Model selection using the maximum log likelihood $\hat{L}$ for each model and AIC.

Model	# params	$\hat{L}$	AIC	ΔAIC
Base	1	15630.6	−31259.2	
Spatial effects (by region)	10	15667.6	−31315.2	−56.0
Spatial effects (by state)	52	15717.2	−31330.4	−71.2
Temporal effects	9	16655.9	−33293.8	−2034.6
Spatial (by region) × temporal effects	90	17290.1	−34400.2	−3141.0
Spatial (by state) × temporal effects	468	18083.6	−35231.2	−3972.0

$$\sigma^{SGTF}_{i,t} = \frac{g^{\Delta H69/V70}_{i,t}}{c_{i,t}\xi^{SGTF}_{i,t}}. \tag{16}$$

Here, $g^{\Delta H69/V70}_{i,t}$ is the number of sequence samples deposited to GISAID with the Δ*H69/V70* deletion.

4.11 Model selection

We initially fit several different models that allowed background transmission rates to vary over space and time in different ways and compared their relative fit using AIC. This model selection step was performed only on the pre-2020-09 data set. Compared to our base model which assumes a constant transmission rate across both space and time, a model that allows transmission rates to vary by geographic region increases the likelihood of the phylogeny and model fit as quantified by AIC (Table 4). A similar model that allowed transmission rates to vary by US state instead of region further improves model fit.

Allowing transmission rates to vary over time in a piecewise constant manner using monthly time intervals improves model fit more than allowing transmission rates to vary by location. Using biweekly rather than monthly time intervals does not improve model fit further. In turn, all models with only spatial or temporal effects are vastly outperformed by a model that allows transmission rates to vary by both time interval and geographic location (spatiotemporal effects). Using states instead of geographic regions increases the likelihood of the spatiotemporal effects model and has lowest overall AIC value, but we continue to use the model with regional spatial resolution as several states are very poorly represented in the GISAID database. We therefore allow transmission rates to vary by geographic region over monthly time intervals in all subsequent analyses.

4.12 Decomposing fitness variation

Given the ancestral features x_n of a lineage, we can compute the lineage's fitness using the fitness mapping function. We can then partition or decompose total variation in fitness between lineages into sources attributable to different components of fitness. To do this, we first partition the features in $\mathcal{X}$ into different disjoint, non-overlapping subsets $\mathcal{X}_k \subset \mathcal{X}$; $\mathcal{X}_k \cap \mathcal{X}_l = \emptyset$ for all subsets k and l.

In the fitness mapping functions presented above, each feature i has a fitness effect $f_{n,i}$ on a lineage's fitness, where $f_{n,i} = \beta_i x_{n,i}$. We let the vector $\boldsymbol{f_i}$ hold the fitness effect of feature i for all lineages in the phylogeny and $\boldsymbol{f}$ hold the overall fitness of each lineage in the phylogeny. Under the additive model that considers fitness on the log scale (2), $\boldsymbol{f} = \sum_i \boldsymbol{f_i}$. Using the general property that the variance in the sum of random variables is equal to the sum of their individual variances and covariances, we can partition the total variation in fitness into variances attributable to individual features and covariances attributable to pairs of features:

$$\mathrm{Var}(\boldsymbol{f}) = \mathrm{Var}\Big(\sum_{i\in\mathcal{X}} \boldsymbol{f_i}\Big) = \sum_{i\in\mathcal{X}} \mathrm{Var}(\boldsymbol{f_i}) + \sum_{i\neq j} \mathrm{Cov}(\boldsymbol{f_i}, \boldsymbol{f_j}) = \sum_{i,j\in\mathcal{X}} \mathrm{Cov}(\boldsymbol{f_i}, \boldsymbol{f_j}), \tag{17}$$

where the covariances account for the fact that the features may be correlated across lineages and therefore not independent.

We can take advantage of the additive property of the variances to compute the fraction of total variance attributable to any particular subset of features $\mathcal{X}_k$:

$$P_k = \frac{\mathrm{Var}\big(\sum_{i\in\mathcal{X}_k} \boldsymbol{f_i}\big)}{\mathrm{Var}\big(\sum_{i\in\mathcal{X}} \boldsymbol{f_i}\big)}. \tag{18}$$

In our SARS-CoV-2 analysis, we partition features into three different components of fitness: genetic, spatial and random (unexplained) effects. To ensure that the fraction of variance attributable to each component sum to one, we compute the fraction of variation attributable to each fitness component as:

$$P_k = \frac{\mathrm{Var}\big(\sum_{i\in\mathcal{X}_k} \boldsymbol{f_i}\big)}{\mathrm{Var}(\boldsymbol{f}_{genetic}) + \mathrm{Var}(\boldsymbol{f}_{spatial}) + \mathrm{Var}(\boldsymbol{f}_{random})}. \tag{19}$$

In other words, we ignore the covariances among fitness components. We do this to ensure that negative covariances among components do not cause the variance attributable to a particular component to be greater than the total variance.

4.13 Two-strain epidemiological model

In order to explore if the transmission fitness effect of approximately 10% that we estimate for Spike D614G would have been sufficient to explain its rapid rise in frequency over the spring of 2020, we simulated the evolutionary dynamics of a mutant variant in a single host population under a two-strain Susceptible-Infected-Exposed-Recovered (SEIR) model parameterized for Covid-19. In this model, an initial resident strain (614G) with transmission rate β seeds the epidemic and then a mutant strain (614D) with transmission rate $\beta_m = \beta f_m$ enters the population through external introductions. We then systematically vary the mutant's fitness f_m to see how much more fit the mutant needs to be in order to match the evolutionary trajectory of D614G.

The epidemic dynamics in the host population are described by the following system of differential equations:

$$
\begin{aligned}
\frac{dS}{dt} &= -S\Big(\beta I + \beta_m I_m\Big) \\
\frac{dE}{dt} &= \beta SI - \eta E \\
\frac{dE_m}{dt} &= \beta_m SI_m - \eta E_m \qquad (20) \\
\frac{dI}{dt} &= \eta E - \nu I + \xi \\
\frac{dI_m}{dt} &= \eta E_m - \nu I_m + \xi_m \\
\frac{dR}{dt} &= \nu\Big(I + I_m\Big).
\end{aligned}
$$

Here, η is the incubation rate at which exposed individuals become infectious and ν is the removal or recovery rate. We assume a 4-day incubation period ($\eta = 0.25$) and a 7-day infectious period ($\nu = 0.143$) (Davies et al., 2020; Ferretti et al., 2020). ξ and ξ_m give the rate of external introductions into the host population (per capita) of the resident and mutant strain, respectively.

We assume that there was a single infected individual in the population on 15 January 2020, reflecting the timing of the earliest probable infections in the US (Worobey et al., 2020). External introductions of the resident strain occur at a rate of 1 per day after Jan. 15th. The external introduction rate of the mutant is initially zero, but switches to $\xi_m > 0$ after 15 February to reflect the earliest probable introductions of D614G into the US. Because the relative rate at which the 614D and 614G entered the US through external introductions is a key unknown that largely determines the evolutionary trajectory of the mutant, we explore different ratios of ξ and ξ_m.

Finally, since no constant transmission rate can recapitulate the epidemic dynamics of Covid-19 in the US under the SEIR model, we allow the base transmission rate β to decline over time to mimic the effects of social distancing or other interventions. We let β decrease between piecewise constant intervals such that R_e is 2.5 between 15 January and 15 February, 1.5 between 15 February and 15 March, 1.25 between 15 March and 15 April and 1.1 after 15 April, reflecting the average R_e values inferred for these time intervals under our phylodynamic model.

Code and data availability

Code and data to replicate our phylodynamic analysis is freely available on GitHub at github.com/davidrasm/phyloTF2.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We would like to thank GISAID and all of the researchers who have submitted SARS-CoV-2 sequences to the GISAID database. A full list of authors and originating laboratories for GISAID submissions we use here is available at: https://github.com/davidrasm/phyloTF2/blob/main/gisaid_hcov-19_acknowledgement_table_2020_12_11_10.pdf. DAR is supported by the US Dept. of Agriculture Hatch project 1016556. LK is supported by the US Centers for Disease Control and Prevention (U01CK000587-01).

Funding

DAR is supported by the US Dept. of Agriculture Hatch project 1016556. LK is supported by the US Centers for Disease Control and Prevention (U01CK000587-01).

Conflict of interest: None declared.

References

Abadi, M. et al. (2016) ‘Tensorflow: A System for Large-Scale Machine Learning’, *12th {USENIX} Symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–83. Savannah, Georgian, USA.

Alizon, S. et al. (2009) ‘Virulence Evolution and the Trade-Off Hypothesis: History, Current State of Affairs and the Future’, *Journal of Evolutionary Biology*, 22: 245–59.

Barido-Sottani, J., Vaughan, T. G., and Stadler, T. (2018) ‘Detection of HIV Transmission Clusters from Phylogenetic Trees using a Multi-State Birth-Death Model’, *Journal of the Royal Society Interface*, 15: 20180512.

Chang, S. et al. (2021) ‘Mobility Network Models of Covid-19 Explain Inequities and Inform Reopening’, *Nature*, 589: 82–7.

Consortium, C. S. M. E. et al. (2004) ‘Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China’, *Science*, 303: 1666–9.

Dalziel, B. D. et al. (2018) ‘Urbanization and Humidity Shape the Intensity of Influenza Epidemics in US Cities’, *Science*, 362: 75–9.

Davies, N. G. et al. (2021) ‘Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England’, *Science*, 372.

——— (2020) ‘Effects of Non-Pharmaceutical Interventions on COVID-19 Cases, Deaths, and Demand for Hospital Services in the UK: a Modelling Study’, *The Lancet Public Health*, 5: e375–85.

Deng, X. et al. (2021) ‘Transmission, Infectivity, and Antibody Neutralization of an Emerging SARS-CoV-2 Variant in California Carrying a L452R Spike Protein Mutation’, *medRxiv*.

Edmonds, C. A., Lillie, A. S., and Cavalli-Sforza, L. L. (2004) ‘Mutations Arising in the Wave Front of an Expanding Population’, *Proceedings of the National Academy of Sciences*, 101: 975–9.

Elbe, S., and Buckland-Merrett, G. (2017) ‘Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health’, *Global Challenges*, 1: 33–46.

Eyre-Walker, A., and Keightley, P. D. (2007) ‘The Distribution of Fitness Effects of New Mutations’, *Nature Reviews Genetics*, 8: 610.

Fauver, J. R. et al. (2020) ‘Coast-to-Coast Spread of SARS-CoV-2 During the Early Epidemic in the United States’, *Cell*, 181: 990–6.

Ferretti, L. et al. (2020) ‘Quantifying SARS-CoV-2 Transmission Suggests Epidemic Control with Digital Contact Tracing’, *Science*, 368: 6491.

Flaxman, S. et al. (2020) ‘Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe’, *Nature*, 584: 257–61.

Foll, M., Shim, H., and Jensen, J. D. (2015) ‘WFABC: a Wright–Fisher ABC-based Approach for Inferring Effective Population Sizes and Selection Coefficients from Time-Sampled Data’, *Molecular Ecology Resources*, 15: 87–98.

Fraser, C. et al. (2007) ‘Variation in HIV-1 Set-Point Viral Load: ePidemiological Analysis and an Evolutionary Hypothesis’, *Proceedings of the National Academy of Sciences*, 104: 17441–6.

Gao, X. et al. (2015) ‘Antibody Against Nucleocapsid Protein Predicts Susceptibility to Human Coronavirus Infection’, *The Journal of Infection*, 71: 599.

Garry, R. F. et al. (2021) 'Spike Protein Mutations in Novel SARS-CoV-2 'Variants of Concern' Commonly Occur in or Near Indels', *virological.org*, 881: 85.

Gillespie, D. T. (2007) 'Stochastic Simulation of Chemical Kinetics', *Annual Review of Physical Chemistry*, 58: 35–55.

Greaney, A. J. et al. (2021) 'Comprehensive Mapping of Mutations to the SARS-CoV-2 Receptor-Binding Domain that Affect Recognition by Polyclonal Human Serum Antibodies', *Cell Host & Microbe*, 29: 463–76.

Hallatschek, O., and Nelson, D. R. (2008) 'Gene Surfing in Expanding Populations', *Theoretical Population Biology*, 73: 158–70.

Handel, A., and Rohani, P. (2015) 'Crossing the Scale from Within-Host Infection Dynamics to Between-Host Transmission Fitness: a Discussion of Current Assumptions and Knowledge', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370: 20140302.

Hodcroft, E. B. et al. (2021) 'Emergence in Late 2020 of Multiple Lineages of SARS-CoV-2 Spike Protein Variants Affecting Amino Acid Position 677', *medRxiv*.

Illingworth, C. J., and Mustonen, V. (2012) 'Components of Selection in the Evolution of the Influenza Virus: Linkage Effects Beat Inherent Selection', *PLoS Pathog*, 8: e1003091.

Ishikawa, S. A. et al. (2019) 'A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios', *Molecular Biology and Evolution*, 36: 2069–85.

Issa, E. et al. (2020) 'SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis', *Msystems*, 5: e00266–20.

Kacser, H., and Burns, J. A. (1981) 'The Molecular Basis of Dominance', *Genetics*, 97: 639–66.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Ke, R. et al. (2020) 'Kinetics of SARS-CoV-2 Infection in the Human Upper and Lower Respiratory Tracts and their Relationship with Infectiousness', *medRxiv*.

Kingma, D. P., and Ba, J. (2014) *Adam: A Method for Stochastic Optimization*, arXiv preprint arXiv:1412.6980.

Kissler, S. et al. (2020) 'Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in New York City', *Nature communications*, 11: 4674.

Klopfstein, S., Currat, M., and Excoffier, L. (2006) 'The Fate of Mutations Surfing on the Wave of a Range Expansion', *Molecular Biology and Evolution*, 23: 482–90.

Korber, B. et al. (2020a) 'Spike Mutation Pipeline Reveals the Emergence of a More Transmissible Form of SARS-CoV-2', *bioRxiv*.

Korber, B. et al. (2020b) 'Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus', *Cell*, 182: 812–27.

Kühnert, D. et al. (2018) 'Quantifying the Fitness Cost of HIV-1 Drug Resistance Mutations through Phylodynamics', *PLoS Pathogens*, 14: e1006895.

Ladner, J. T. et al. (2021) 'Epitope-Resolved Profiling of the SARS-CoV-2 Antibody Response Identifies Cross-Reactivity with Endemic Human Coronaviruses', *Cell Reports Medicine*, 2: 100189.

Lanfear, R. (2020). *A Global Phylogeny of hcov-19 Sequences from Gisaid*.

Larsen, B. B., and Worobey, M. (2021) 'Phylogenetic Evidence that B.1.1.7 has been Circulating in the United States since Early- to Mid-November', *virological.org*. <https://virological.org/t/phylogenetic-evidence-that-b-1-1-7-has-been-circulating-in-the-united-states-since-early-to-mid-november/598> accessed 30 Aug 2021.

Lei, J., Kusov, Y., and Hilgenfeld, R. (2018) 'Nsp3 of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein', *Antiviral Research*, 149: 58–74.

Leung, N. H. et al. (2020) 'Respiratory Virus Shedding in Exhaled Breath and Efficacy of Face Masks', *Nature Medicine*, 26: 676–80.

Lloyd-Smith, J. O. et al. (2005) 'Superspreading and the Effect of Individual Variation on Disease Emergence', *Nature*, 438: 355–9.

Long, J. S. et al. (2016) 'Species Difference in ANP32A Underlies Influenza A Virus Polymerase Host Restriction', *Nature*, 529: 101–4.

MacLean, O. A. et al. (2020a) 'Natural Selection in the Evolution of SARS-CoV-2 in Bats, Not Humans, Created a Highly Capable Human Pathogen', *BioRxiv*.

MacLean, O. A. et al. (2020b) 'No Evidence for Distinct Types in the Evolution of SARS-CoV-2', *Virus Evolution*, 6: veaa034.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007) 'Estimating a Binary Character's Effect on Speciation and Extinction', *Systematic Biology*, 56: 701–10.

Maliet, O., Hartig, F., and Morlon, H. (2019) 'A Model with Many Small Shifts for Estimating Species-Specific Diversification Rates', *Nature Ecology and Evolution*, 3: 1086–92.

Martin, D. P. et al. (2021) 'The Emergence and Ongoing Convergent Evolution of the N501Y Lineages Coincides with a Major Global Shift in the SARS-CoV-2 Selective Landscape', *medRxiv*.

Muth, D. et al. (2018) 'Attenuation of Replication by a 29 Nucleotide Deletion in SARS-Coronavirus Acquired During the Early Stages of Human-to-Human Transmission', *Scientific Reports*, 8: 1–11.

Nadeau, S. A. et al. (2021) 'The Origin and Early Spread of SARS-CoV-2 in Europe', *Proceedings of the National Academy of Sciences*, 118.

Naveca, F. et al. (2021) 'Phylogenetic Relationship of SARS-CoV-2 Sequences from Amazonas with Emerging Brazilian Variants Harboring Mutations E484K and N501Y in the Spike Protein', *Virological.org*. <https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585> accessed 30 Aug 2021.

Neher, R. A. (2013) 'Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation', *Annual Review of Ecology, Evolution, and Systematics*, 44: 195–215.

Neher, R. A., Russell, C. A., and Shraiman, B.I. (2014) 'Predicting Evolution from the Shape of Genealogical Trees', *Elife*, 3: e03568.

Pater, A. A. et al. (2021) 'Emergence and Evolution of a Prevalent New SARS-CoV-2 Variant in the United States', *bioRxiv*.

Perez-Saez, J. et al. (2021) 'Serology-Informed Estimates of SARS-CoV-2 Infection Fatality Risk in Geneva, Switzerland', *The Lancet Infectious Diseases*, 21: e69–70.

Plante, J. A. et al. (2021) 'Spike Mutation D614G Alters SARS-CoV-2 Fitness', *Nature*, 592: 116–21.

Ragonnet-Cronin, M. et al. (2021) 'Genetic Evidence for the Association between COVID-19 Epidemic Severity and Timing of Non-Pharmaceutical Interventions', *Nature Communications*, 12: 1–7.

Rambaut, A. et al. (2020) 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.

Rasigade, J.-P. et al. (2020). *A Viral Perspective on Worldwide Non-Pharmaceutical Interventions against COVID-19*.

Rasmussen, D. A., and Stadler, T. (2019) 'Coupling Adaptive Molecular Evolution to Phylodynamics Using Fitness-Dependent Birth-Death Models', *eLife*, 8: e45562.

Roberts, D. R. et al. (2017) 'Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure', *Ecography*, 40: 913–29.

Sanjuán, R., Moya, A., and Elena, S. F. (2004) 'The Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in an RNA Virus', *Proceedings of the National Academy of Sciences*, 101: 8396–401.

Sankoff, D. (1975) 'Minimal Mutation Trees of Sequences', *SIAM Journal on Applied Mathematics*, 28: 35–42.

Shaman, J., and Kohn, M. (2009) 'Absolute Humidity Modulates Influenza Survival, Transmission, and Seasonality', *Proceedings of the National Academy of Sciences*, 106: 3243–8.

Stadler, T. (2009) 'On Incomplete Sampling Under Birth–Death Models and Connections to the Sampling-Based Coalescent', *Journal of Theoretical Biology*, 261: 58–66.

Stadler, T., and Bonhoeffer, S. (2013) 'Uncovering Epidemiological Dynamics in Heterogeneous Host Populations Using Phylogenetic Methods', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368: 20120198.

Stadler, T. et al. (2013) 'Birth–Death Skyline Plot Reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (HCV)', *Proceedings of the National Academy of Sciences*, 110: 228–33.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Starr, T. N. et al. (2020) 'Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding', *Cell*, 182: 1295–310.

Sukumaran, J., and Holder, M. T. (2010) 'DendroPy: a Python Library for Phylogenetic Computing', *Bioinformatics*, 26: 1569–71.

Tang, X. et al. (2020) 'On the Origin and Continuing Evolution of SARS-CoV-2', *National Science Review*, 7: 1012–23.

Tegally, H. et al. (2020) 'Emergence and Rapid Spread of a New Severe Acute Respiratory Syndrome-Related Coronavirus 2 (SARS-CoV-2) Lineage with Multiple Spike Mutations in South Africa', *medRxiv*.

To, T.-H. et al. (2015) 'Fast Dating Using Least-Squares Criteria and Algorithms', *Systematic Biology*, 65: 82–97.

Tuffley, C., and Steel, M. (1997) 'Links between Maximum Likelihood and Maximum Parsimony Under a Simple Model of Site Substitution', *Bulletin of Mathematical Biology*, 59: 581–607.

Urbanowicz, R. A. et al. (2016) 'Human Adaptation of Ebola Virus During the West African Outbreak', *Cell*, 167: 1079–87.

Volz, E. et al. (2021) 'Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity', *Cell*, 184: 64–75.

Volz, E. et al. (2021) 'Transmission of SARS-CoV-2 Lineage B. 1.1. 7 in England: Insights from Linking Epidemiological and Genetic Data', *medRxiv*, 2020–12.

Walensky, R. P., Walke, H. T., and Fauci, A. S. (2021) 'SARS-CoV-2 Variants of Concern in the United States—Challenges and Opportunities', *JAMA*, 325: 1037–8.

Washington, N. L. et al. (2021) 'Genomic Epidemiology Identifies Emergence and Rapid Transmission of SARS-CoV-2 B. 1.1.7 in the United States', *medRxiv*.

Washington, N. L. et al. (2020) 'S Gene Dropout Patterns in SARS-CoV-2 Tests Suggest Spread of the H69del/V70del Mutation in the US', *medRxiv*.

Wölfel, R. et al. (2020) 'Virological Assessment of Hospitalized Patients with COVID-2019', *Nature*, 581: 465–9.

Worobey, M. et al. (2020) 'The Emergence of SARS-CoV-2 in Europe and North America', *Science*, 370: 564–70.

Wu, S. L. et al. (2020) 'Substantial Underestimation of SARS-CoV-2 Infection in the United States', *Nature Communications*, 11: 1–10.

Xue, K. S., and Bloom, J. D. (2020) 'Linking Influenza Virus Evolution within and between Human Hosts', *Virus Evolution*, 6: veaa010.

Zahradnik, J. et al. (2021) 'SARS-CoV-2 RBD in Vitro Evolution Follows Contagious Mutation Spread, yet Generates an able Infection Inhibitor', *BioRxiv*.

Zhang, J., and Nei, M. (1997) 'Accuracies of Ancestral Amino Acid Sequences Inferred by the Parsimony, Likelihood, and Distance Methods', *Journal of Molecular Evolution*, 44: S139–46.

Zhang, L. et al. (2020) 'SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity', *Nature Communications*, 11: 1–9.

Zhang, W. et al. (2021) 'Emergence of a Novel SARS-CoV-2 Variant in Southern California', *JAMA*, 325: 1324–6.