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

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

1 Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA
2 Biomathematics Graduate Program, North Carolina State University, Raleigh, North Carolina, USA
3 Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, North Carolina, USA

* Contributed equally to this work
E-mail: drasmus@ncsu.edu

December 13, 2020

Abstract

The fitness of a pathogen is 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 behaviour influence between-host transmission potential. These challenges have 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 and September 2020 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. Up to September 2020, most fitness variation between lineages can be explained by background spatial heterogeneity in transmission rates across geographic regions. Furthermore, no genetic variant including the Spike D614G mutation has had a significant effect on population-level fitness. Instead, the rapid increase in the frequency of the Spike D614G can be explained by the variant having a spatial transmission advantage due to first establishing in regions with higher transmission rates during the earliest stages of the pandemic.

Introduction

Determining what factors shape the overall fitness of a novel pathogen such as SARS-CoV-2 is a 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 received considerable attention (Korber et al., 2020a; MacLean et al., 2020; Tang et al., 2020). The most notable of these variants is 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 outcompeted other viral genotypes that were already established in several locations (Korber et al., 2020b). Further evidence for a fitness benefit came from experimental studies showing that the D614G variant increases cellular infectivity and viral replication rates both in vitro and in vivo (Korber et al., 2020b; Plante et al., 2020; Zhang et al., 2020). However, the role of D614G in driving the overall epidemic and its impact on population-level fitness remain disputed (Grubaugh et al., 2020), with estimates of the fitness effect of the D614G variant ranging from low to moderately large benefits (Leung et al., 2020a; Volz et al., 2020). Other mutations, including several structural variants containing large deletions, have been reported to affect virulence or disease outcomes in clinical settings (Liu et al., 2020; Su et al., 2020; Young et al., 2020), but the fitness effects of these mutations at the population-level have not been explored.

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 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., 2020b). Thus, in order to provide a definitive answer about the epidemiological significance of a novel pathogen variant, fitness also needs to be quantified at the between-host or population-level.

Fitness at the population level can be inferred based on the evolutionary dynamics of pathogen variants in the host population. For example, the growth rate of alternate variants can be estimated from time series of variant frequencies or pathogen phylogenies as a surrogate for fitness (Foll et al., 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 pathogen fitness while controlling for multiple confounding factors. On average, a pathogen lineage with increased population-level 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 et al., 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 et al., 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 determine the fitness of SARS-CoV-2 at the host population-level in the United States. 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 pathogen 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.

Results

Phylogenetic and ancestral state reconstruction

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 October 2nd, 2020. Dated or time-calibrated maximum likelihood phylogenetic trees were then reconstructed from whole genome sequences. For all sampled sequences, we also assembled a set of features that potentially predict fitness, including both genetic and non-genetic, environmental features. The genetic features include 66 amino acid and 6 structural (deletion) variants that were present in at least 0.5% of all sequences sampled in the Unites States up to September 1st. 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.

Background spatial and temporal effects

Before considering models that include genetic variants as fitness-predicting features, we consider several models accounting for background spatial and temporal variability in transmission, which could otherwise confound fitness estimates. 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 1). 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 piece-wise 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.

Our maximum likelihood estimate (MLE) of the base transmission rate across all times and regions is 0.16 per day. Assuming a constant recovery/removal rate of 0.14 per day yields an estimate of the basic

Table 1. 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) x temporal effects	90	17290.1	-34400.2	-3141.0
Spatial (by state) x temporal effects	468	18083.6	-35231.2	-3972.0

reproduction number $R_0 = 1.15$. This base R_0 estimate is considerably lower than previous estimates which generally range from 2.0 - 3.0 (Li et al., 2020; Ke et al., 2020a), but the spatiotemporal effects rescale the base transmission rate by region and time, providing a range of effective reproduction numbers R_e between 1.05 and 5.84 (Figure 1).

As expected, estimated transmission rates and R_e vary dramatically across the US by both region and time, with New York and New Jersey (Region 2), the Upper Midwest (Region 5) and the Pacific Northwest (Region 10) having the highest transmission at the beginning of the pandemic in early February and other regions peaking slightly later (Figure 1). Transmission rates estimated from the phylogeny peak in the period between February 1st and March 1st, substantially earlier than peaks in reported cases. This pattern has been reported in other phylodynamic studies (Fauver et al., 2020; Nadeau et al., 2020; Ragonnet-Cronin et al., 2020), and may reflect considerable undetected transmission as well as lags in reporting before routine testing began. Estimated transmission rates then remain low through spring and early summer but rise again in late summer across all regions.

So far we have assumed a constant sampling fraction ($s = 0.0004$, see Methods) and not accounted for differences in sampling or sequencing effort across time and space. While it is theoretically possible to estimate both transmission rates and sampling fractions from phylogenies (Stadler, 2009), these two parameters are generally highly negatively correlated such that jointly estimating both is not possible without making additional assumptions (Louca and Pennell, 2020). In order to ensure that our estimates were not unduly influenced by sampling biases, we estimated how sampling fractions varied across space and time by considering the number of sequences submitted to GISAID relative to the total number of SARS-CoV-2 cases imputed from reported Covid-related deaths in each region (see Methods). Although imputed sampling fractions vary by several orders of magnitude (Supp. Figure 1), explicitly accounting for sampling biases in this way did not significantly alter estimated transmission rates (Supp. Figure 2). Thus, whether or not we explicitly consider sampling biases, we can account for spatiotemporal background heterogeneity in transmission rates (or sampling) and thereby control for their confounding effects when estimating the fitness effects of genetic variants.

Fitness effects of genetic variants

We next estimated the fitness effect of genetic variants while controlling for background heterogeneity in transmission rates using the spatial-temporal effects model. We consider the fitness effect of 66 amino acid variants in coding regions spanning the SARS-CoV-2 genome, however several variants are tightly linked and nearly always co-occur together (Supp. Figure 3), leading to strong collinearity among features in our model. We therefore encode sets of linked variants that are over 95% correlated with one another as single features.

Fitness effects were estimated under a model where each variant has a multiplicative effect on the overall fitness of a lineage such that a neutral variant has a fitness of 1.0 and deleterious or beneficial mutants have fitness effects less than or greater than 1.0, respectively. We only consider a variant to be significantly deleterious or beneficial if the 95% credible interval (CI) does not overlap with 1.0.

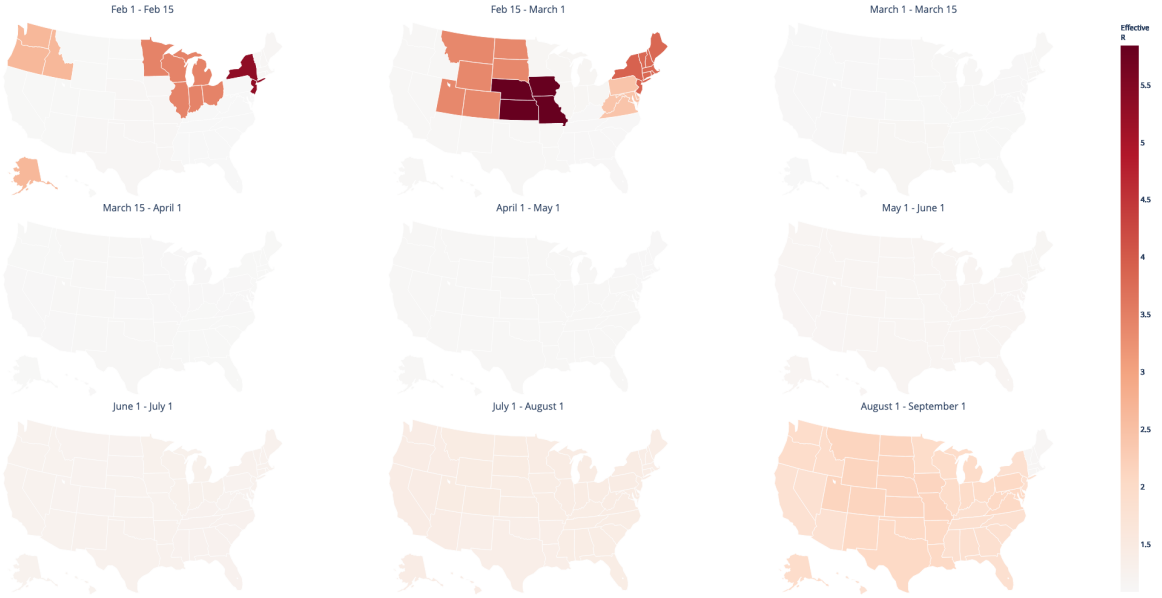


Figure 1. Background spatiotemporal heterogeneity in the effective reproductive number R_e of SARS-CoV-2 inferred from the ML viral phylogeny. A regional transmission effect was estimated for each region and time interval and then used to rescale our base estimate of the transmission rate to compute R_e. Note that here we use biweekly rather than monthly time intervals at the beginning of the pandemic to more clearly illustrate spatiotemporal heterogeneity among regions. States are grouped into the geographic regions designated by the US Department of Health and Human Services.

Most amino acid variants and linked variant sets are inferred to be neutral, with maximum likelihood estimates of fitness effects close to 1.0 and 95% CIs overlapping 1.0 (Figure 2). In addition to the best ML phylogeny, we estimated fitness effects from 10 bootstrapped phylogenies on which ancestral features were independently reconstructed. MLEs of fitness effects are highly consistent across bootstrap replicates.

Only one set of linked variants, nsp13 L504P+C541Y, is estimated to be significantly but weakly deleterious with a fitness effect of 0.99 (95% CI: 0.98-0.991). However, it is worth noting that there is likely an ascertainment bias against the inclusion of strongly deleterious mutations, as these variants would likely not have reached a frequency above our inclusion criteria of 0.5%. Two variants were estimated to have significantly positive but minor fitness effects, nsp13 E224D (MLE: 1.02; 95% CI: 1.01-1.03) and ORF9 S190I (MLE: 1.033; 95% CI: 1.02-1.05). There were also two pairs of linked variants with significantly positive fitness effects, nsp1 D144A+ORF9 S235F (MLE: 1.025; 95% CI: 1.01-1.04) and Spike F1052L+ORF6 L4P (MLE: 1.031; 95% CI: 1.01-1.05).

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. Despite rapidly increasing in frequency in the spring of 2020 (Figure 3A), the Spike D614G + nsp12 P323L variant is estimated to have a only a modest fitness benefit of 1.011 (95% CI: 1.00-1.013).

In addition to amino acid variants, we considered the fitness effects of several structural variants, including large deletions in nsp14, ORF7a, ORF7b, ORF8 and the transcriptional regulatory sequence (TRS) of ORFS 7b and 8. While there were several other structural variants circulating, these deletion mutations were among the most frequent and deemed most likely to be functionally important. nsp14 encodes an exoribonuclease (ExoN) involved in RNA proofreading which likely plays a critical role in

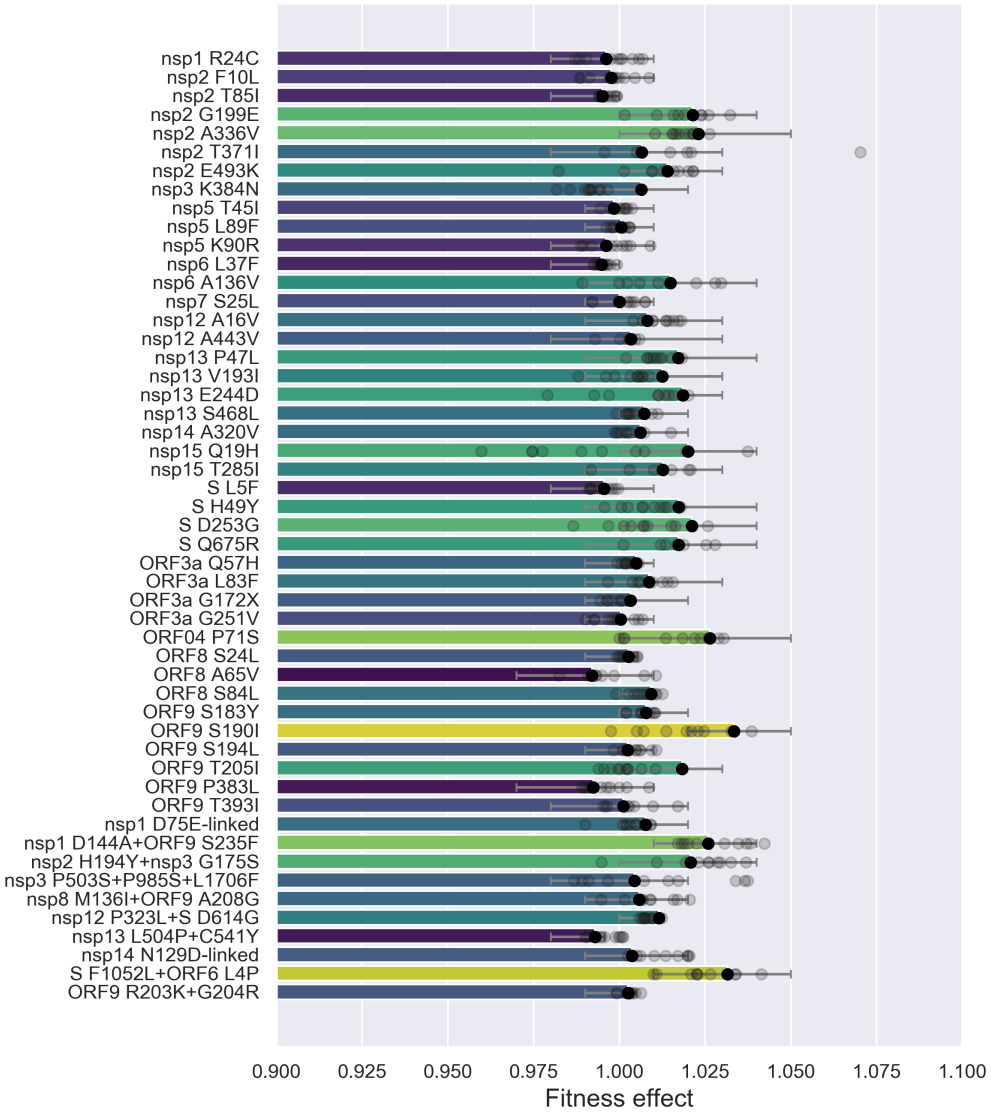


Figure 2. Estimated fitness effects of amino acid variants. Fitness effects are jointly estimated under a model of multiplicative fitness, such that neutral variants have a fitness of one. Variants are ordered from top to bottom by their genomic position. Bars are colored according to the maximum likelihood estimate (MLE) of each variant's fitness effect. Capped lines indicate the 95% credible interval around the MLE. The MLE of each fitness effect is also shown for 10 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.

replication fidelity (Snijder et al., 2003; Eckerle et al., 2007; Smith et al., 2013), although the deletion considered here occurs in the C-terminal N7-MT domain required for viral mRNA cap synthesis (Chen

et al., 2013). ORFs 7 and 8 are hot spots of major deletion mutants in the coronavirus genome and encode for accessory proteins that mediate the host-immune system (Guan et al., 2003; Su et al., 2020; Young et al., 2020). Nevertheless, all structural variants were estimated to have a nearly neutral fitness effect with credible intervals overlapping 1.0 (Supp. Figure 4).

Explaining the dominance 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 614 variants themselves. Lineages with the 614G variant have an average background fitness that is 3.8% higher than the 614D variant. After partitioning total background fitness into genetic and spatial components, the 614G variant occurs in genetic backgrounds with 0.7% higher fitness and spatial backgrounds (i.e. geographic regions) with 2.9% higher fitness on average. However, these averages conceal the fact that the background fitness advantage of the 614G variant derived almost entirely from being in geographic regions with higher average transmission rates during the earliest stages of the pandemic (Figure 3). 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 (Supp. Figure 5).

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 (Supp. Figure 6). 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 (Supp. Figure 7). To place the variants on more equitable footing, we therefore compare the branching/transmission rate of the variants *per lineage*, which accounts for the 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 (Figure 4), supporting our model-based inference that the 614G variant has no or only a very minor 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 = 0.53 per week) is slightly higher than the 614D variant (mean = 0.46 per week), but these means are not significantly different (Welch's t-test = -1.42; p-value = 0.15).

Decomposing the sources of fitness variation

We next 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

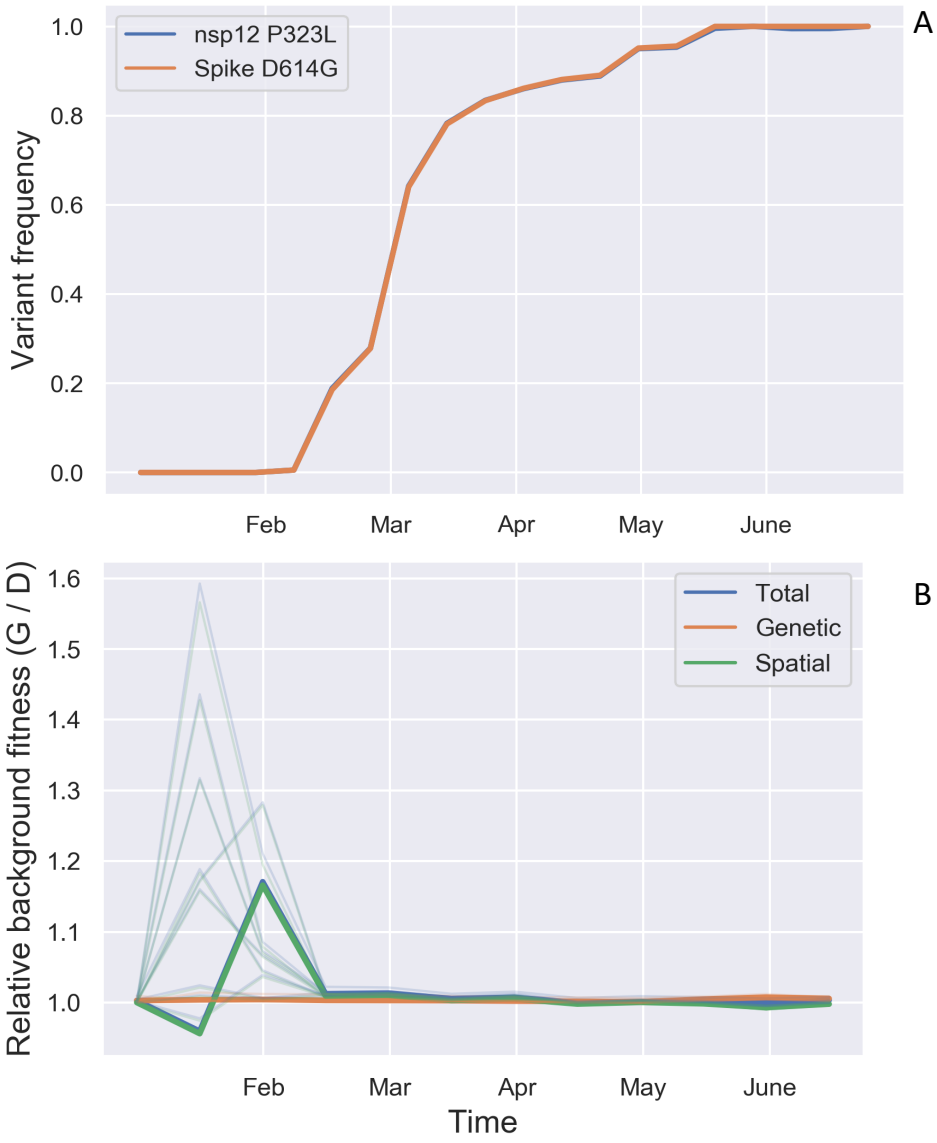


Figure 3. Evolutionary dynamics and background fitness of the Spike 614 variants. (A) Frequency of lineages carrying the Spike D614G and nsp12 P323L variants over time relative to all lineages in the ML phylogeny. These two variants are tightly linked so that they largely share the same evolutionary 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. The analysis was performed on the best ML phylogeny (solid lines) and 10 replicate bootstrapped trees (transparent lines). Relative fitness is only shown up to July 1st, 2020 as the 614D variant was not sampled after this date.

that we can use to predict the relative fitness of each lineage. Fitness varies considerably between lineages in the phylogeny as well as over time (Figure 5).

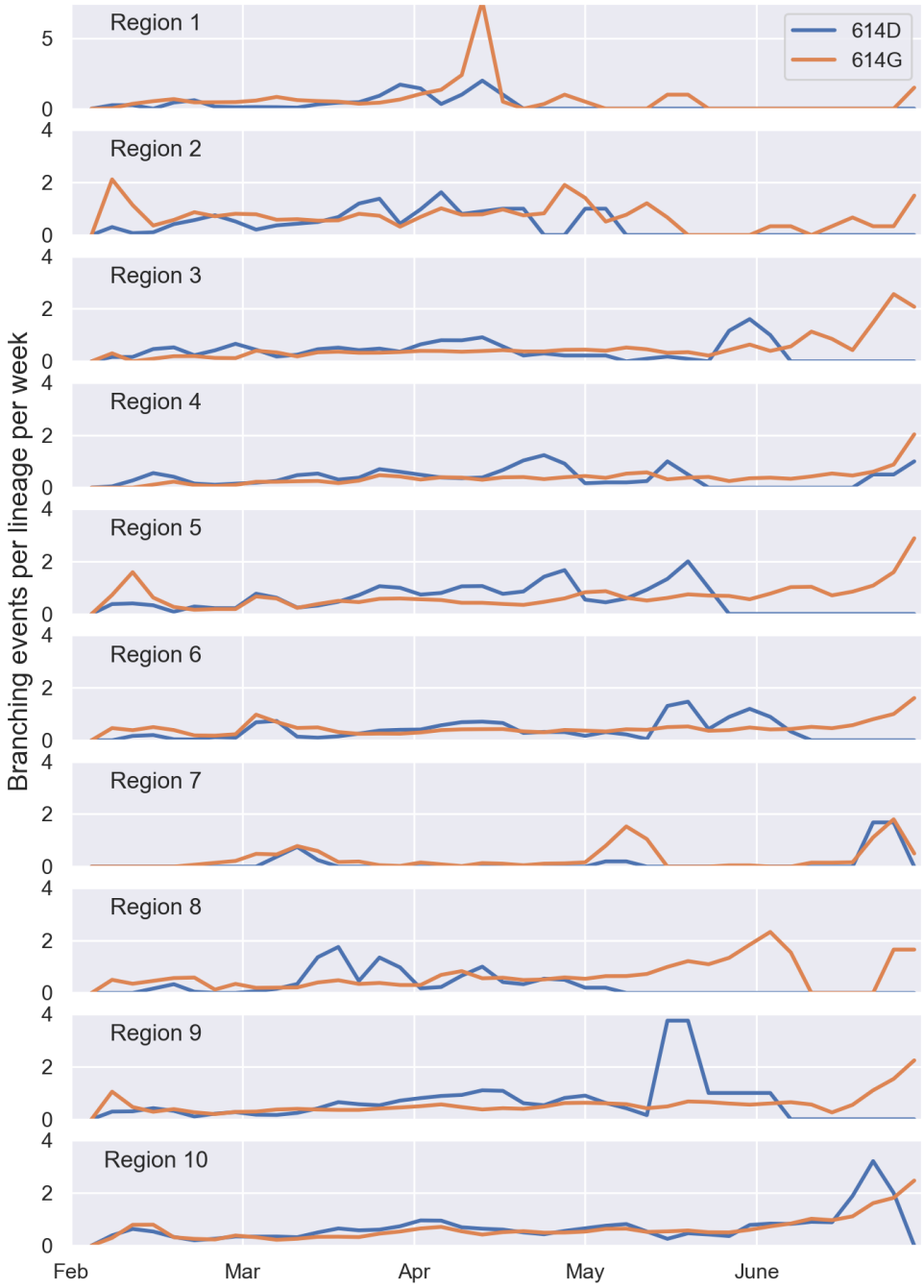


Figure 4. Branching rates per lineage in each region for lineages with the Spike 614D versus 614G variant. Branching rates are reported here as the number of branching events per week in the ML phylogeny for lineages with either variant.

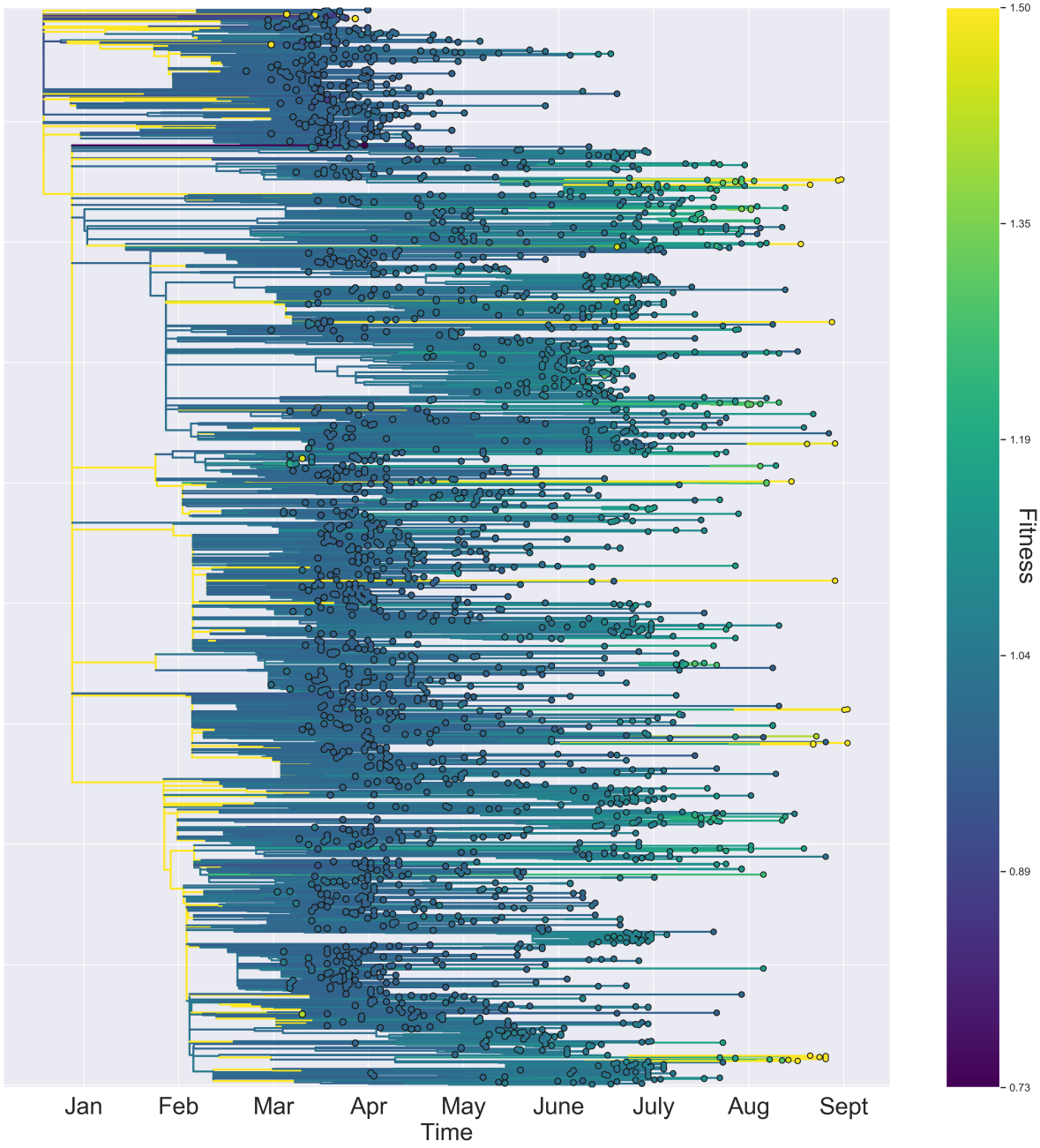


Figure 5. SARS-CoV-2 phylogeny with lineages colored by their predicted relative fitness. Fitness is predicted based on each lineage's ancestral features using the fitted fitness mapping function with spatial, genetic and random effects. The tree was thinned to include only 10% of all sampled tips in the full tree for purposes of visualization. The range of the color map was also capped at a value of 1.5 to emphasize variation in fitness surrounding the mean rather than the full range of fitness values.

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 (Figure 6). On average, about 50% of total fitness variation is attributable to random effects (Figure 6), which in turn implies that 50% of fitness variation is explained by spatial effects or genetic features in our model. At the beginning of the pandemic, virtually all fitness variation among lineages is attributable to spatial heterogeneity in transmission among geographic regions, but the contribution of spatial effects declines over time and then rises again in late summer. As expected, genetic variants contribute no fitness variation at the beginning of the pandemic when the virus population was genetically homogeneous. However, an increasing fraction of fitness variation is attributable to genetic variants over time and up to 25% of all fitness variation was explained by genetic variants in early summer, although total fitness variation in the population was low during this time period. Note that genetic variation in fitness cannot be explained by the Spike 614 variants as the 614D variant was already extremely rare by early summer. Rather genetic fitness variation appears to be due to a large number of variants each with relatively small fitness effects circulating in the population.

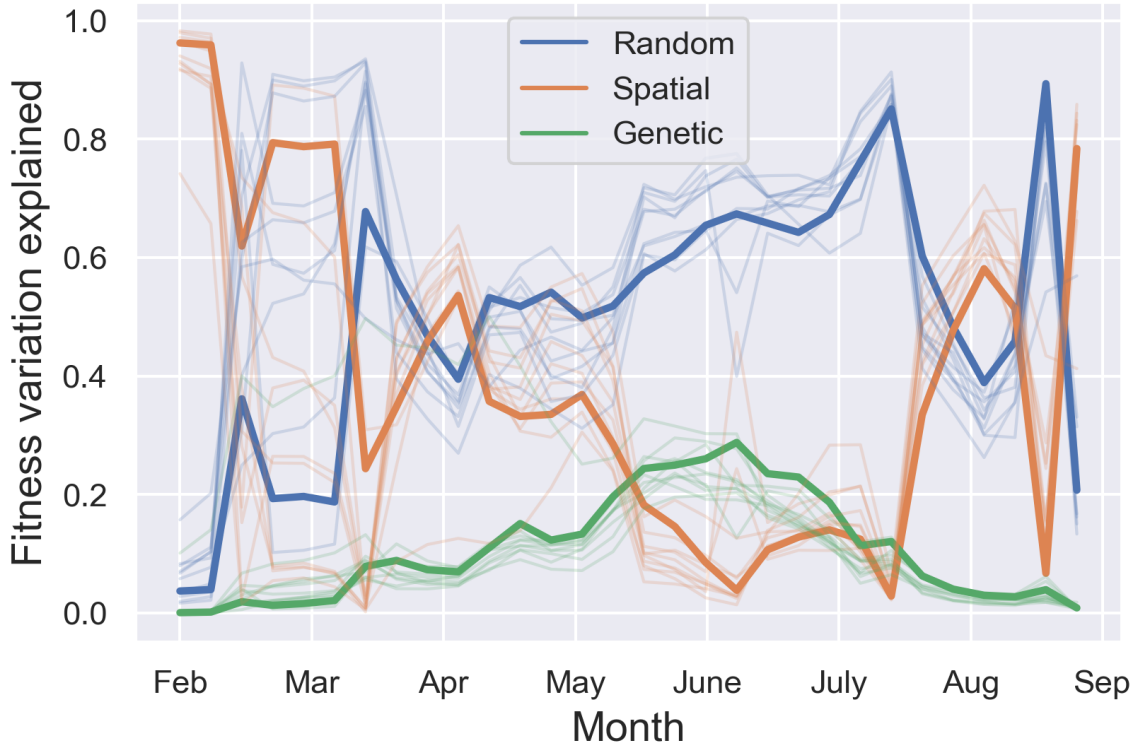


Figure 6. Fitness variation among lineages decomposed into sources attributable to different components of fitness. All lineages present in the phylogeny at a give time point were used to compute fitness variation among lineages. The variance decomposition was performed on the best ML phylogeny (solid lines) and 10 replicate bootstrapped trees (transparent lines).

Discussion

We 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. Applying this framework to a data set including over 22,000 viruses sampled in the United States over the first nine months of the pandemic, our results suggest that fitness variation among lineages is largely driven by non-genetic or extrinsic factors. Phylodynamic estimates indicate that transmission rates varied considerably between geographic regions, especially early in the pandemic, and a large fraction of total fitness variation is attributable to spatial heterogeneity in transmission. In contrast no genetic variant, including the Spike D614G mutation, was determined to have a major impact on fitness. We conclude that up to this point in time genetic variants have contributed little to overall fitness variation in the SARS-CoV-2 viral population.

What extrinsic factors underlie the spatial or temporal variability in transmission rates? 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., 2020) 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 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 behaviour 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 (Rasigade et al., 2020).

While all of the amino acid and structural variants we considered here are estimated to have near neutral fitness effects at the population-level, there is growing evidence that variants such as the Spike D614G mutation can significantly alter viral fitness within individual hosts. Controlled experiments show that the D614G mutation allows the Spike glycoprotein to increase its binding affinity to the human ACE2 receptor, increasing cellular infectivity and viral replication rates in both cell culture and hamsters (Korber et al., 2020b; Plante et al., 2020). 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., 2020). We do not intend to cast doubt on these experimental results, but stress 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 (Ke et al., 2020b; Wölfel et al., 2020). Instead infectiousness increases with the logarithm of viral load such that it saturates at higher viral loads. 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., 2020b). 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. Furthermore, recent deep mutational scanning experiments indicate that the binding affinity of the ancestral 614D variant to the ACE2 receptor may be perfectly adequate, as there are accessible mutations in Spike receptor binding domain that increase ACE2 binding affinity even further than circulating mutations but these mutations are not selected for in cell culture experiments (Starr et al., 2020). The enhanced cellular infectivity and replication rates of D614G within hosts is therefore not irreconcilable with our inferences that the mutant has a very modest if any population-level fitness effect.

It does initially seem more challenging to reconcile the D614G mutation having no strong selective advantage with its rapid spread and near universal rise in frequency around the world. Our phylodynamic

estimates are also at odds with estimates from time series of variant frequencies, which indicate that the 614G variant has a 20-30% selective advantage based on its rapid increase in frequency (Leung et al., 2020a; Volz et al., 2020). Nevertheless, the recent phylodynamic analysis of Volz et al. (2020) in the UK and our own analysis in the US consistently estimate either no or only a minor fitness advantage for the 614G variant. Our phylogenetic analysis may partially explain these disparate estimates. 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 in the US. This scenario is strongly supported by our phylodynamic analysis, which revealed that lineages carrying the 614G variant had a higher average background fitness due to occurring in regions with higher transmission rates rather than any intrinsic or genetic fitness advantage. While the 614G variant subsequently increased to high frequencies in all geographic regions, this rapid increase can partially be explained by more lineages carrying the 614G variant being imported into each region (Supp. Figure 7). A rapid increase in the frequency of the 614G variant over the 614D variant therefore does not imply that the 614G variant had any inherent transmission advantage. In fact, when comparing the relative branching rate of lineages with either variant as a surrogate for transmission rates, we find no consistent difference in their branching rates in any region.

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 et al., 2004; Klopfstein et al., 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 prospective, 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.

While our phylodynamic inference framework accounts for many potentially confounding factors including background fitness variation, 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. Although we did not try to directly correct for sampling biases in the GISAID database, we included spatiotemporal effects in our model in order to account for differences in either background transmission rates or sampling fractions over time. Moreover, estimated transmission rates did not significantly vary depending on whether we assume a constant sampling fraction or try to explicitly model how sampling fractions vary over space and time. Second, we chose a simple fitness mapping function that assumes each feature has a multiplicative effect on lineage fitness (such that log fitness is an additive linear function of features). 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 (GxE interactions). Learning what types of functions are expressive enough to capture these these complexities while remaining statistically tractable and biologically interpretable is a major challenge for future work. Finally, the computational efficiency of our approach relies on first reconstructing phylogenies and ancestral states before fitting our phylodynamic birth-death model. While we partially account 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. New inference methods are clearly needed to fit complex phylodynamic models to genomic data sets as large as those currently available for SARS-CoV-2.

Even if our present analysis strongly suggests that up to this point in time genetic variants have only played a minor role in shaping the fitness of SARS-CoV-2, the situation may change rapidly in the future. For example, as natural or vaccine-induced immunity builds in the human population, antigenic mutations may arise that allow the virus to escape immunity. In this case, our phylodynamic framework could 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 that up to now seem to explain most fitness variation. 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.

Models and Methods

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 et al., 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.

Phylogenetic reconstruction

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 October 2nd, 2020 representing sequences that were sequenced prior to September 1st, 2020. Genomes were aligned using MAFFT (Katoh and Standley, 2013). A maximum likelihood (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.

Ancestral state reconstruction

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 each lineage n, ancestral features were then combined into a vector of categorical variables x_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.

Fitness mapping functions

Our main goal is to learn the fitness mapping function $F(x_n)$ that maps the features of a lineage x_n to that lineage's expected fitness. While $F(x_n)$ could be any arbitrary function, we use a simple model that assumes the fitness effect β_i of each feature i is multiplicative:

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

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:

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

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

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

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 birth/transmission 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 et al. (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.

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 x_n using a fitness mapping function $F(x_n)$. 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 further 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:

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

The likelihood of an individual branching or transmission event is:

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

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

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

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

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 et al. (2018), over a

time interval of length Δ_t, this likelihood can be computed as:

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

with:

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

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

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

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:

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

$E(0)$ 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.

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:

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

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 maximum likelihood estimates (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% credible intervals using an asymptotic chi-square approximation to the likelihood ratio test.

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 x 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 (Supp. Figure 8). 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.

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, 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.

We also 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 assume $\sigma = 0.0004$, although we consider a model below where σ is allowed to vary over time and by region. This estimate was back-calculated based on the number of cumulative deaths that had occurred in the US relative to the number of US samples in the GISAID database. On July 16th 2020, the day we first downloaded sequence data, there had been an estimated 130,371 cumulative deaths in the US (The COVID Project: `https://covidtracking.com/data/national`). Assuming an overall mortality rate of 0.5% gives 26,074,200 cumulative cases by July 16th. Our initial July 16th sequence data set included 10,483 samples, providing a crude estimate of σ = 10,483 / 26,074,200 = 0.0004.

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 piece-wise constant and add each lineage segment to the set of lineages in N.

Modeling sampling heterogeneity

In the model presented above, we implicitly account for sampling heterogeneity by allowing the transmission rate to vary across space and time but keep the sampling fraction σ fixed at a constant rate. Our estimated transmission rates will therefore likely be biased due to ignoring sampling heterogeneity, but by allowing the estimated transmission rates to vary across space and time we hope to account for background heterogeneity in either transmission or sampling that could otherwise confound estimates of fitness effects of genetic variants.

To further ensure that ignoring sampling heterogeneity does not significantly bias our estimates, we consider another model where we explicitly track how sampling fractions vary across space and time. In this model, 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 cases 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 case fatality ratio ϕ (assumed to be 0.5% as above). 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 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 (Supp. Figure 1). For these estimates we assume reported deaths lag behind reported cases by three weeks when estimating sampling fractions. At the beginning of the epidemic there is extreme geographic heterogeneity in sampling fractions which vary across regions by over four orders of magnitude, but by late summer the sampling fraction across the country converges on a value close to the national average assumed above (0.004).

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:

$$\mathrm{Var}(\boldsymbol{f}) = \mathrm{Var}\Big(\sum_{i\in\mathcal{X}} \boldsymbol{f_i}\Big) = \sum_{i\in\mathcal{X}} \mathrm{Var}(\boldsymbol{f_i}) + \sum_{i\neq j} \mathrm{Cov}(\boldsymbol{f_i}, \boldsymbol{f_j}) = \sum_{i,j\in\mathcal{X}} \mathrm{Cov}(\boldsymbol{f_i}, \boldsymbol{f_j}), \tag{16}$$

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 $\mathcal{X}_k$:

$$P_k = \frac{\mathrm{Var}\big(\sum_{i\in\mathcal{X}_k} X_i\big)}{\mathrm{Var}\big(\sum_{i\in\mathcal{X}} X_i\big)} \tag{17}$$

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:

$$P_k = \frac{\mathrm{Var}\big(\sum_{i\in\mathcal{X}_k} X_i\big)}{\mathrm{Var}(f_{genetic}) + \mathrm{Var}(f_{spatial}) + \mathrm{Var}(f_{random})} \tag{18}$$

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.

Code and data availability

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

Acknowledgments

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).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Alizon, S., Hurford, A., Mideo, N., and Van Baalen, M. (2009). Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *Journal of Evolutionary Biology*, 22(2):245–259.

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(146):20180512.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. (2020). Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, pages 1–8.

Chen, Y., Tao, J., Sun, Y., Wu, A., Su, C., Gao, G., Cai, H., Qiu, S., Wu, Y., Ahola, T., et al. (2013). Structure-function analysis of severe acute respiratory syndrome coronavirus RNA cap guanine-N7-methyltransferase. *Journal of Virology*, 87(11):6296–6305.

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(5664):1666–1669.

Dalziel, B. D., Kissler, S., Gog, J. R., Viboud, C., Bjørnstad, O. N., Metcalf, C. J. E., and Grenfell, B. T. (2018). Urbanization and humidity shape the intensity of influenza epidemics in us cities. *Science*, 362(6410):75–79.

Eckerle, L. D., Lu, X., Sperry, S. M., Choi, L., and Denison, M. R. (2007). High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *Journal of Virology*, 81(22):12135–12144.

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(4):975–979.

Elbe, S. and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46.

Fauver, J. R., Petrone, M. E., Hodcroft, E. B., Shioda, K., Ehrlich, H. Y., Watts, A. G., Vogels, C. B., Brito, A. F., Alpert, T., Muyombwe, A., et al. (2020). Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell*.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261.

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(1):87–98.

Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F., and Hanage, W. P. (2007). Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences*, 104(44):17441–17446.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55.

Grubaugh, N. D., Hanage, W. P., and Rasmussen, A. L. (2020). Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell*, 182(4):794–795.

Guan, Y., Zheng, B., He, Y., Liu, X., Zhuang, Z., Cheung, C., Luo, S., Li, P., Zhang, L., Guan, Y., et al. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*, 302(5643):276–278.

Hallatschek, O. and Nelson, D. R. (2008). Gene surfing in expanding populations. *Theoretical Population Biology*, 73(1):158–170.

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(1675):20140302.

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(12):e1003091.

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A fast likelihood method to reconstruct and visualize ancestral scenarios. *Molecular Biology and Evolution*, 36(9):2069–2085.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Ke, R., Romero-Severson, E. O., Sanche, S., and Hengartner, N. (2020a). Estimating the reproductive number R0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. *medRxiv*.

Ke, R., Zitzmann, C., Ribeiro, R. M., and Perelson, A. S. (2020b). 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., Kishore, N., Prabhu, M., Goffman, D., Beilin, Y., Landau, R., Gyamfi-Bannerman, C., Bateman, B., Katz, D., Gal, J., et al. (2020). Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City.

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(3):482–490.

Korber, B., Fischer, W., Gnanakaran, S. G., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E. E., Bhattacharya, T., Parker, M. D., et al. (2020a). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., et al. (2020b). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182(4):812–827.

Kühnert, D., Kouyos, R., Shirreff, G., Pečerska, J., Scherrer, A. U., Böni, J., Yerly, S., Klimkait, T., Aubert, V., Günthard, H. F., et al. (2018). Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. *PLoS Pathogens*, 14(2):e1006895.

Leung, K., Pei, Y., Leung, G. M., Lam, T. T., and Wu, J. T. (2020a). Empirical transmission advantage of the D614G mutant strain of SARS-CoV-2. *medRxiv*.

Leung, N. H., Chu, D. K., Shiu, E. Y., Chan, K.-H., McDevitt, J. J., Hau, B. J., Yen, H.-L., Li, Y., Ip, D. K., Peiris, J. M., et al. (2020b). Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nature Medicine*, 26(5):676–680.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490):489–493.

Liu, Z., Zheng, H., Lin, H., Li, M., Yuan, R., Peng, J., Xiong, Q., Sun, J., Li, B., Wu, J., et al. (2020). Identification of common deletions in the spike protein of SARS-CoV-2. *Journal of Virology*.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

Long, J. S., Giotis, E. S., Moncorgé, O., Frise, R., Mistry, B., James, J., Morisson, M., Iqbal, M., Vignal, A., Skinner, M. A., et al. (2016). Species difference in ANP32A underlies influenza A virus polymerase host restriction. *Nature*, 529(7584):101–104.

Louca, S. and Pennell, M. W. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505.

MacLean, O. A., Orton, R. J., Singer, J. B., and Robertson, D. L. (2020). No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evolution*, 6(1):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(5):701–710.

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(7):1086–1092.

Muth, D., Corman, V. M., Roth, H., Binger, T., Dijkman, R., Gottula, L. T., Gloza-Rausch, F., Balboni, A., Battilani, M., Rihtarič, 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):1–11.

Nadeau, S. A., Vaughan, T. G., Sciré, J., Huisman, J. S., and Stadler, T. (2020). The origin and early spread of SARS-CoV-2 in Europe. *medRxiv*.

Neher, R. A. (2013). Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):195–215.

Neher, R. A., Russell, C. A., and Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *Elife*, 3:e03568.

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., et al. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, pages 1–9.

Ragonnet-Cronin, M., Boyd, O., Geidelberg, L., Jorgensen, D., Nascimento, F. F., Siveroni, I., Johnson, R., Baguelin, M., Cucunuba, Z. M., Jauneikaite, E., et al. (2020). Covid-19 epidemic severity is associated with timing of non-pharmaceutical interventions. *medRxiv*.

Rasigade, J.-P., Barray, A., Shapiro, J. T., Coquisart, C., Vigouroux, Y., Bal, A., Destras, G., Vanhems, P., Lina, B., Josset, L., 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., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Shaman, J. and Kohn, M. (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9):3243–3248.

Smith, E. C., Blanc, H., Vignuzzi, M., and Denison, M. R. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. *PLoS Pathog*, 9(8):e1003565.

Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., Guan, Y., Rozanov, M., Spaan, W. J., and Gorbalenya, A. E. (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *Journal of molecular biology*, 331(5):991–1004.

Stadler, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):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(1614):20120198.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.

Su, Y. C., Anderson, D. E., Young, B. E., Linster, M., Zhu, F., Jayakumar, J., Zhuang, Y., Kalimuddin, S., Low, J. G., Tan, C. W., et al. (2020). Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. *MBio*, 11(4).

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65(1):82–97.

Urbanowicz, R. A., McClure, C. P., Sakuntabhai, A., Sall, A. A., Kobinger, G., Müller, M. A., Holmes, E. C., Rey, F. A., Simon-Loriere, E., and Ball, J. K. (2016). Human adaptation of Ebola virus during the West African outbreak. *Cell*, 167(4):1079–1087.

Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., et al. (2020). Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell*.

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809):465–469.

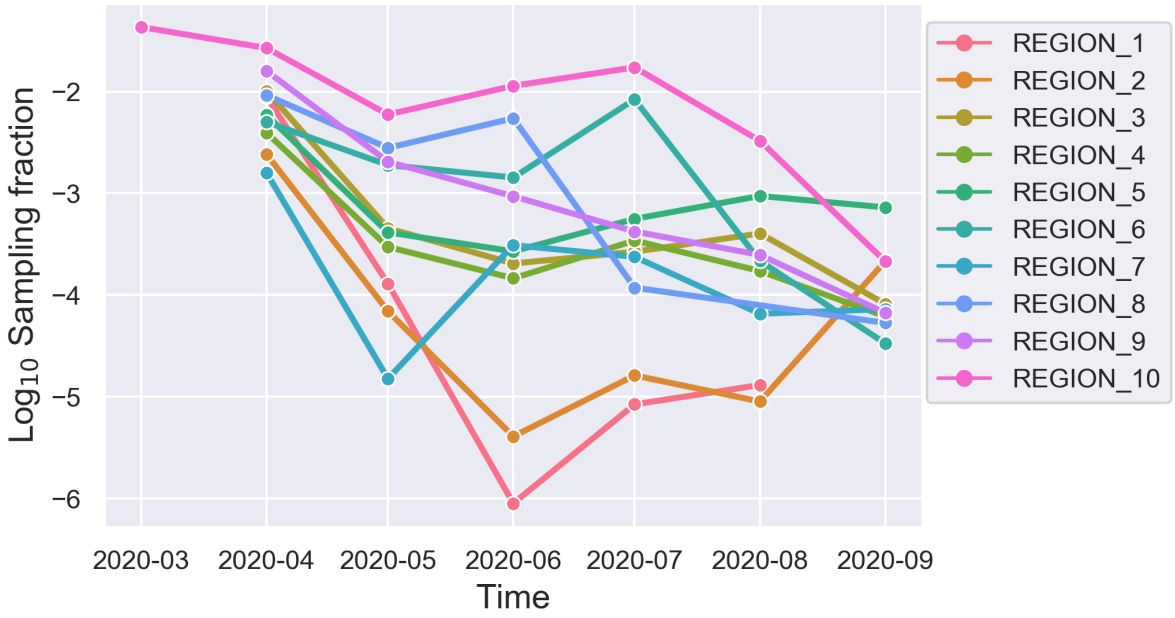
Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., et al. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, 11(1):1–10.

Xue, K. S. and Bloom, J. D. (2020). Linking influenza virus evolution within and between human hosts. *Virus Evolution*, 6(1):veaa010.

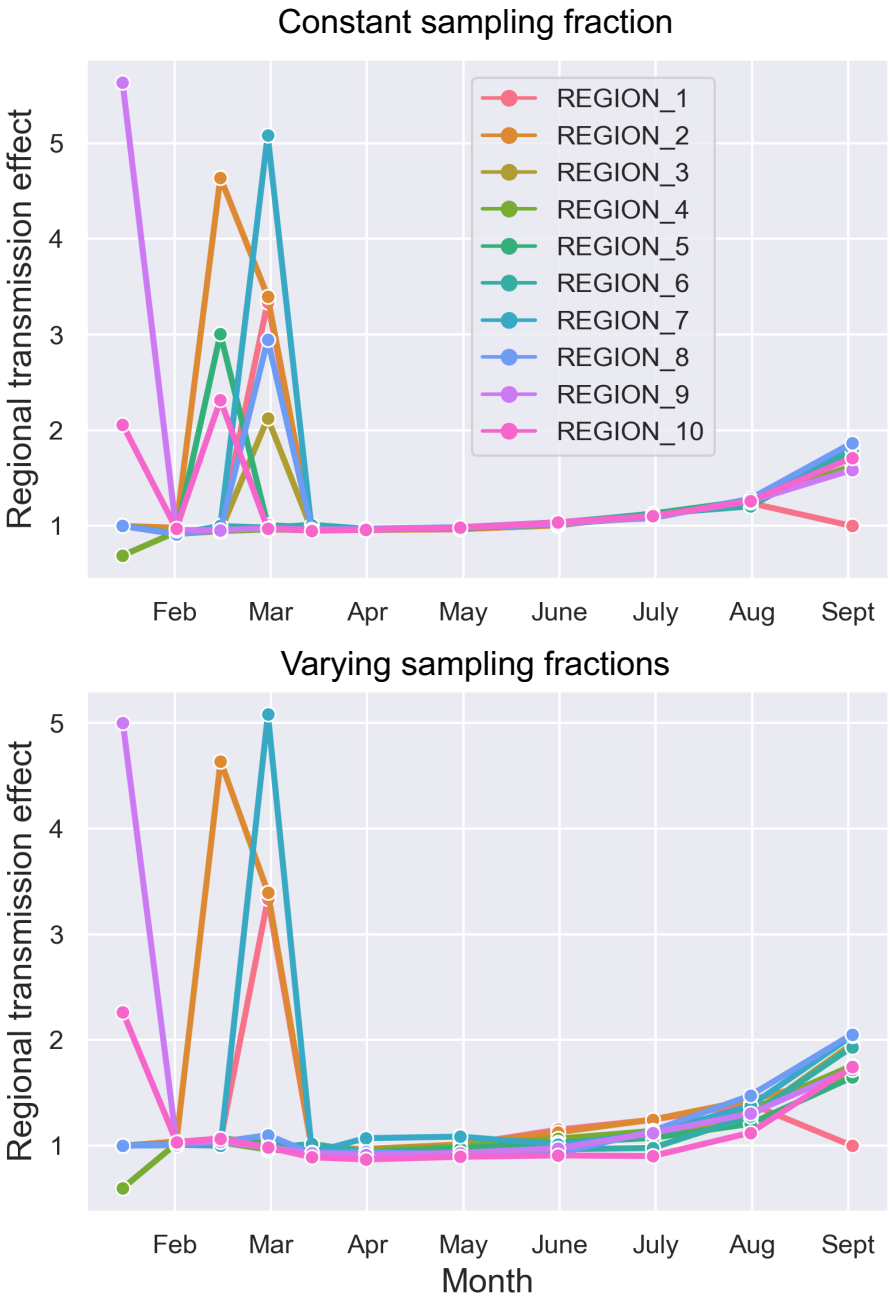
Young, B. E., Fong, S.-W., Chan, Y.-H., Mak, T.-M., Ang, L. W., Anderson, D. E., Lee, C. Y.-P., Amrun, S. N., Lee, B., Goh, Y. S., et al. (2020). Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *The Lancet*.

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., Farzan, M., and Choe, H. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv*.

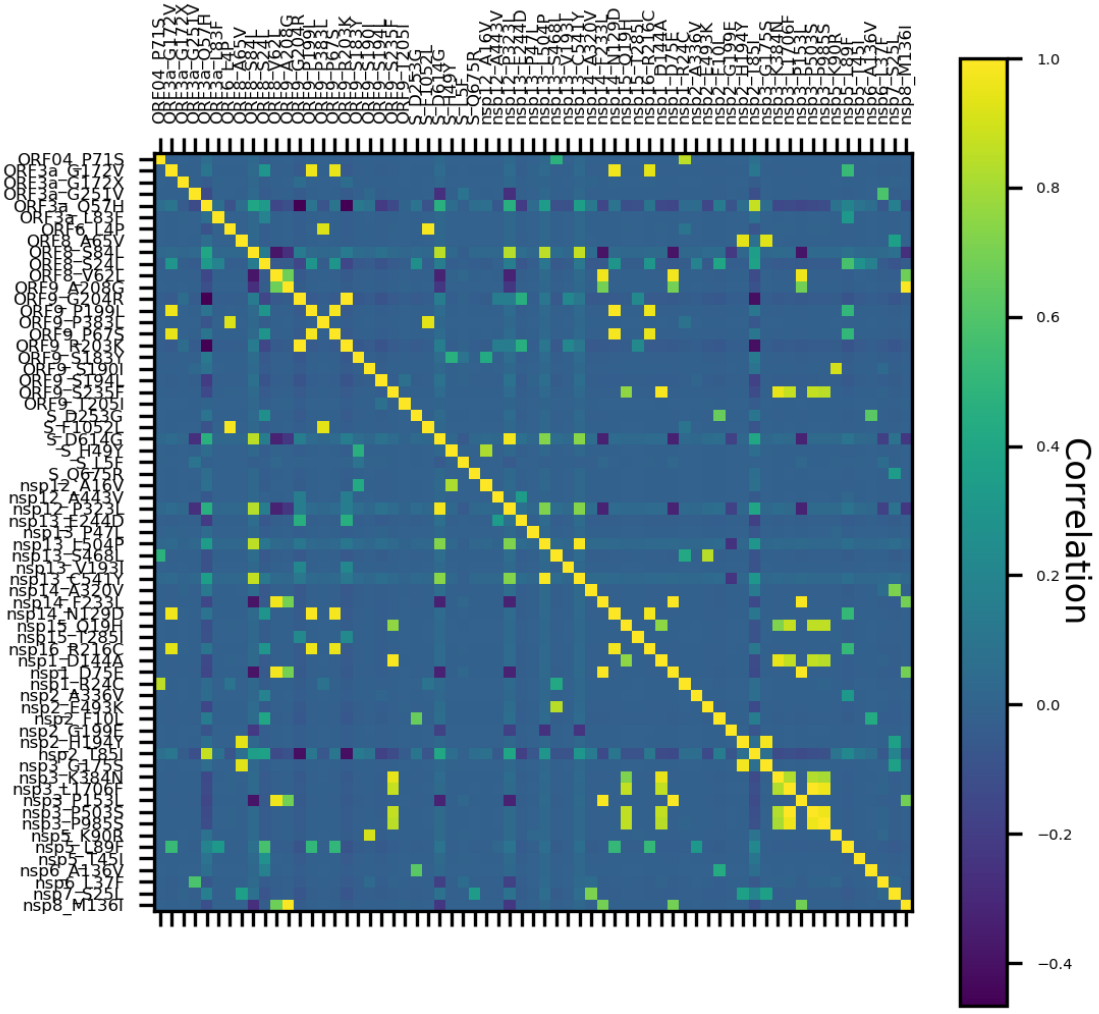
Supplementary Figures



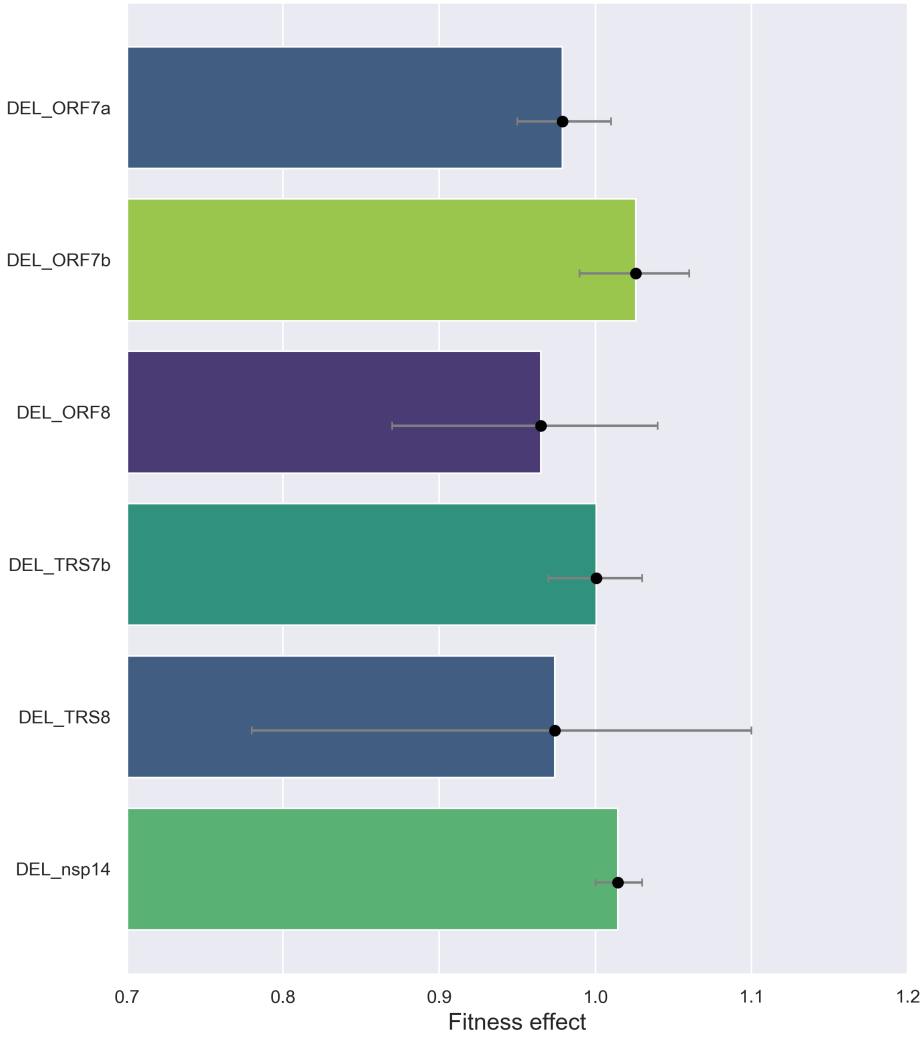
Supp. Figure 1. Sampling fractions based on the number of sequences deposited in GISAID. Sampling fractions were estimated by dividing the number of samples in the GISAID database by the estimated number of total cases for each region and time interval.



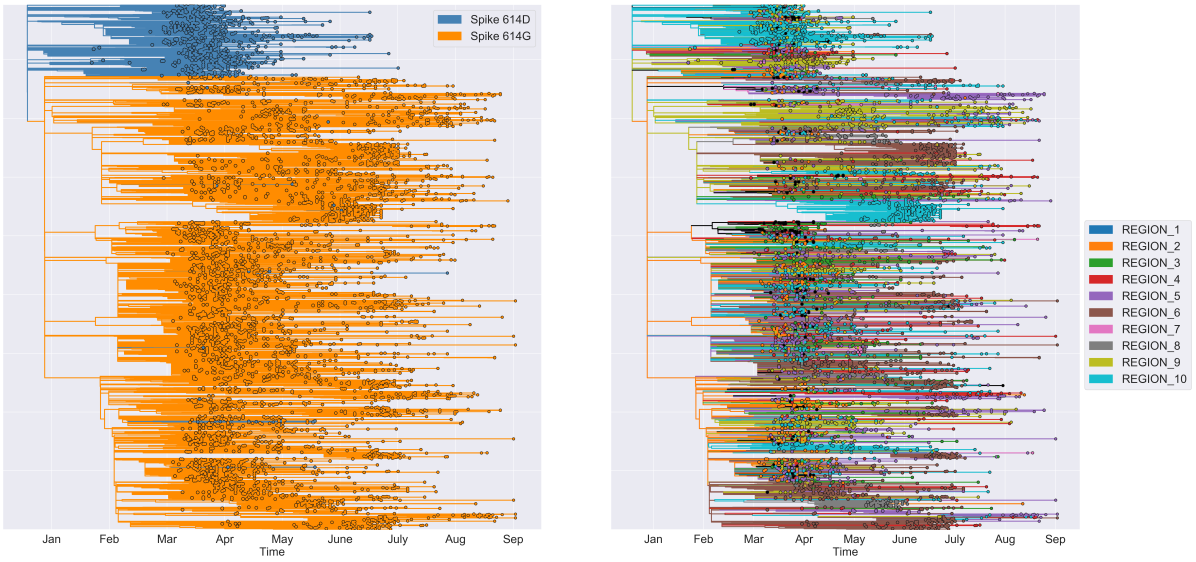
Supp. Figure 2. Background spatiotemporal heterogeneity in SARS-CoV-2 transmission. A regional transmission effect was estimated for each region over one month intervals. In the top row, transmission effects were estimated assuming a constant sampling fraction ($s = 0.0004$) across space and time. In the bottom row, sampling fractions were allowed to vary over space and time using using the estimates shown in Supp. Figure 1.



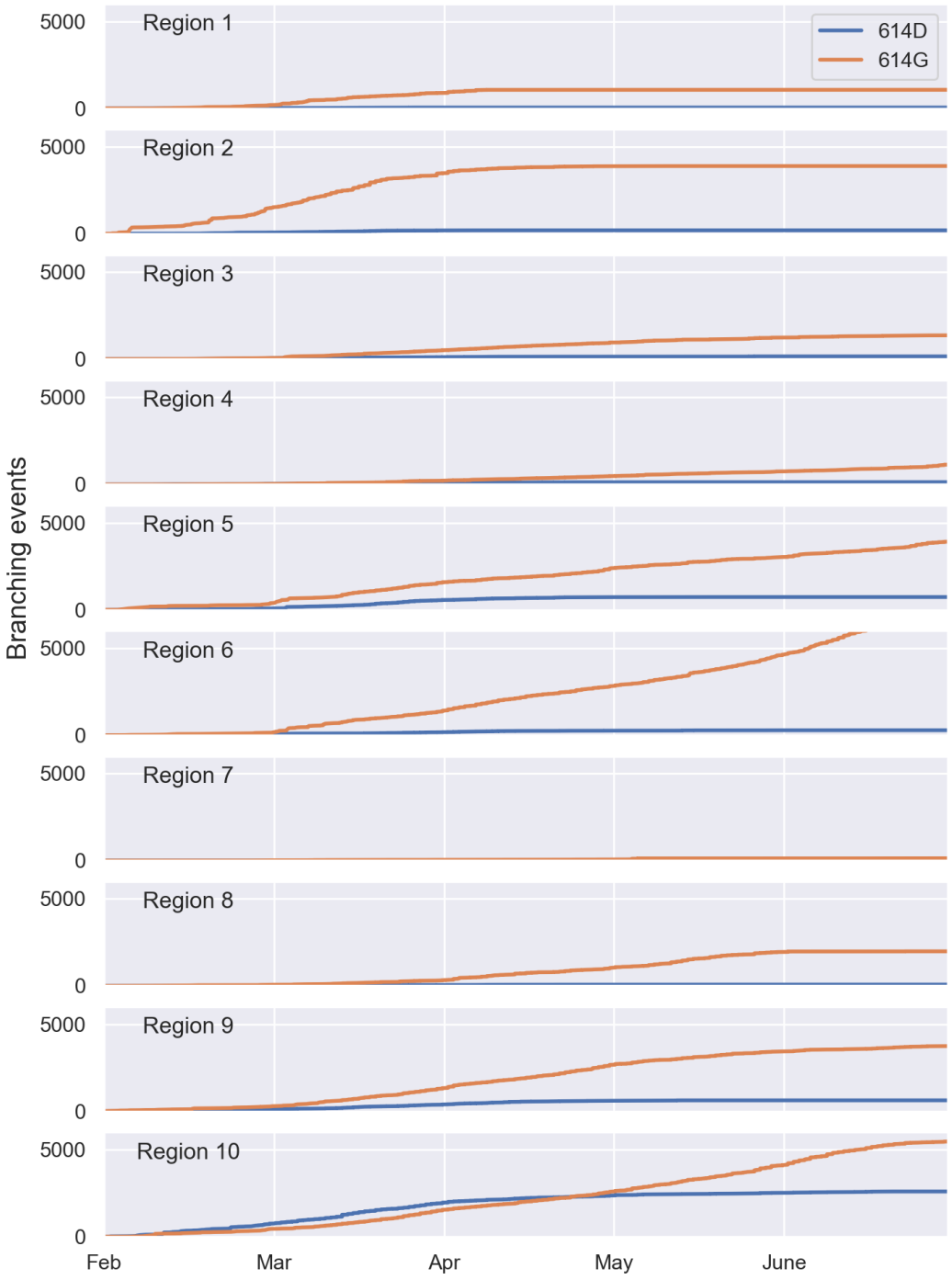
Supp. Figure 3. Correlation among amino acid variants in sampled SARS-CoV-2 genotypes. Colors in the heat map represent the Pearson correlation coefficient between each pair of variants.



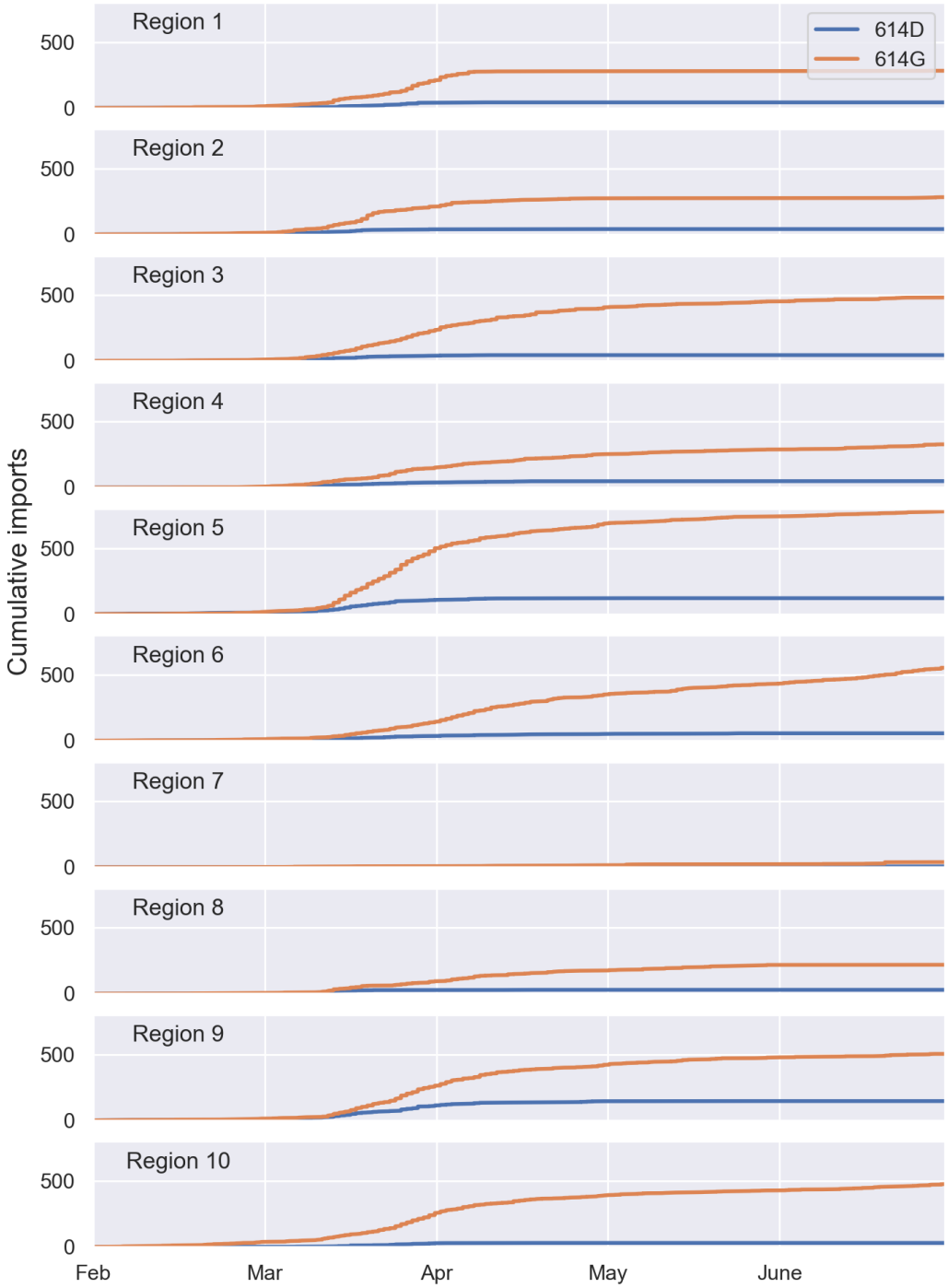
Supp. Figure 4. Estimated fitness effects of structural variants with major deletions. Bars are colored according to the maximum likelihood estimate of the fitness effect of each variant. Capped lines indicate 95% credible intervals. Deletion mutations in these regions tend to be hypervariable in both number and their starting and end positions. We therefore grouped all structural variants with major deletions in one of these six regions. In the reference Wuhan-Hu-1 SARS-CoV-2 genome, these deletions correspond to positions: 27397-27736 (DEL ORF7a), 27759-27891 (DEL ORF7b), 27897-28262 (DEL ORF8), 27708-27752 (DEL TRS7b), 27884-27897(DEL TRS8) and 19276-19579 (DEL nsp14).



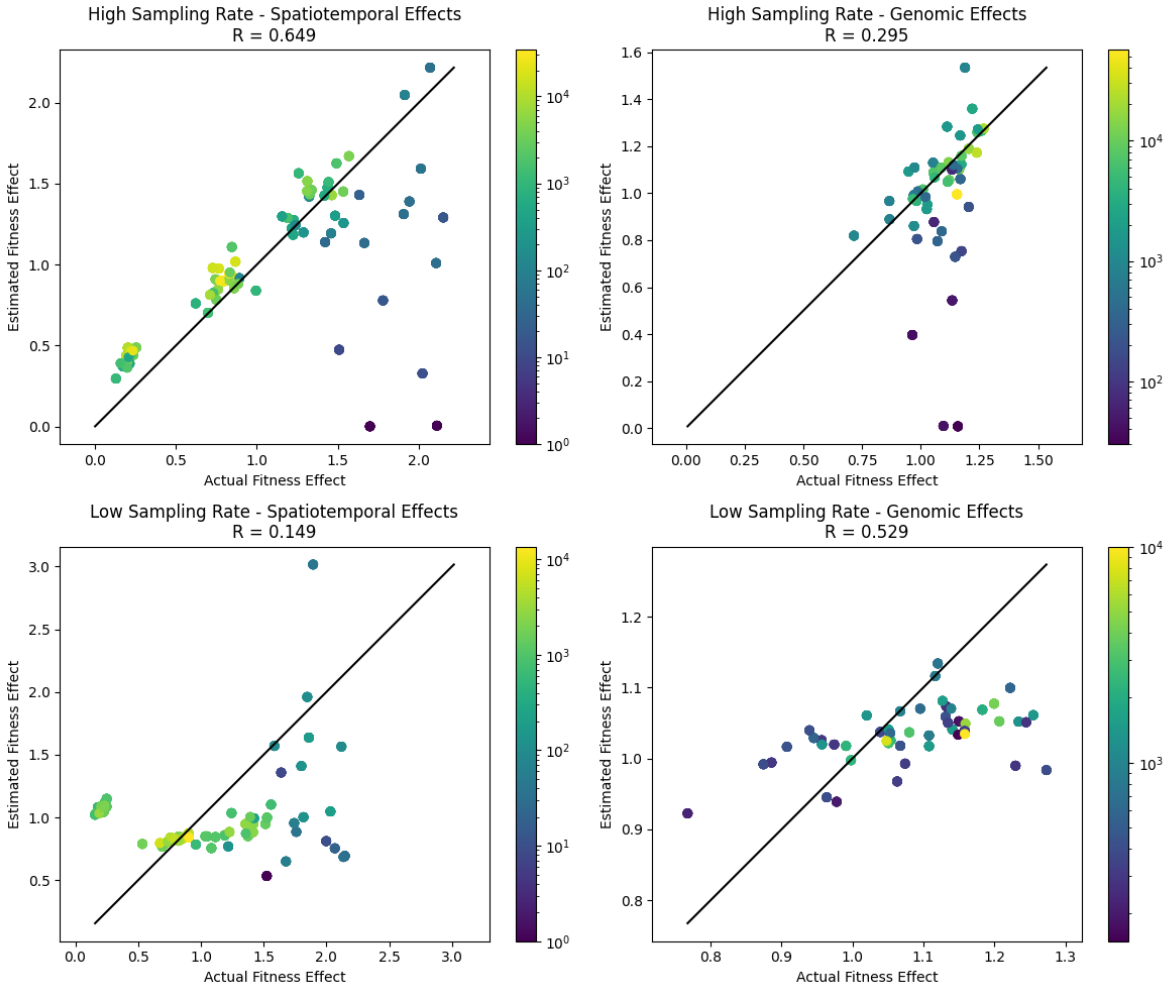
Supp. Figure 5. Maximum likelihood SARS-CoV-2 phylogenies with reconstructed ancestral features. The reconstructed Spike 614 variant type is shown on the left . The reconstructed geographic location in terms of US HHS Regions is shown on the left. The trees were thinned to only include 10% of all sampled tips in the full tree for purposes of visualization.



Supp. Figure 6. Cumulative branching events along lineages with the Spike 614 G versus D variant. Branching events are grouped by region based on their reconstructed ancestral location at the branching node.



Supp. Figure 7. Cumulative number of lineages imported into each region with the Spike 614 G or D variant. Importation events were identified based on the reconstructed ancestral location of lineages in the phylogeny and defined here as a lineage's ancestral location changing between a parent and a child node. The age/height of the child node was taken to be the time of the introduction event.



Supp. Figure 8. Actual (true) versus estimated fitness effects for features inferred from simulated phylogenetic trees. Dots represent the actual and estimated fitness value for an individual feature and are colored according to the number of individuals sampled with the corresponding feature. True fitness effects were known from simulations under a stochastic birth-death-sampling model with mutation. In these simulations, spatiotemporal (regional) and genomic (mutational) fitness effects were drawn randomly for each feature. In the top row sampling fractions are high ($\sigma = \rho = 0.5$) whereas in the bottom row sampling fractions are low ($\sigma = \rho = 0.05$). Simulation results were pooled across 10 simulations in each plot. Fitness estimates for features with a sampling fraction of less than 0.5% were discarded.