PAPER

Inferring selection effects in SARS-CoV-2 with Bayesian Viral Allele Selection

Martin Jankowiak,[1,*] Fritz H. Obermeyer[2,†] and Jacob E. Lemieux[1,3]

[1]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA, [2]Generate Biomedicines, Cambridge, Massachusetts, USA and [3]Division of Infectious Diseases, Massachusetts General Hospital, Boston, Massachusetts, USA
*Corresponding author: mjankowi@broadinstitute.org †Work done at the Broad Institute.

Abstract

The global effort to sequence millions of SARS-CoV-2 genomes has provided an unprecedented view of viral evolution. Characterizing how selection acts on SARS-CoV-2 is critical to developing effective, long-lasting vaccines and other treatments, but the scale and complexity of genomic surveillance data make rigorous analysis challenging. To meet this challenge, we develop Bayesian Viral Allele Selection (BVAS), a principled and scalable probabilistic method for inferring the genetic determinants of differential viral fitness and the relative growth rates of viral lineages, including newly emergent lineages. After demonstrating the accuracy and efficacy of our method through simulation, we apply BVAS to 6.9 million SARS-CoV-2 genomes. We identify numerous mutations that increase fitness, including previously identified mutations in the SARS-CoV-2 Spike and Nucleocapsid proteins, as well as mutations in non-structural proteins whose contribution to fitness is less well characterized. In addition, we extend our baseline model to identify mutations whose fitness exhibits strong dependence on vaccination status as well as pairwise interaction effects, i.e. epistasis. Strikingly, both these analyses point to the pivotal role played by the N501 residue in the Spike protein. Our method, which couples Bayesian variable selection with a diffusion approximation in allele frequency space, lays a foundation for identifying fitness-associated mutations under the assumption that most alleles are neutral.

Key words: SARS-CoV-2, Bayesian Methods, Viral Epidemiology, GWAS, Diffusion Processes

1. Introduction

The SARS-CoV-2 pandemic has seen the repeated emergence of new viral lineages with higher fitness, where fitness includes any attribute that affects the lineage's growth, including its basic reproduction number and generation time. The virus has evolved into numerous sublineages that are characterized by distinct phenotypes including enhanced pathogenicity, increased escape from convalescent and vaccine-acquired immunity, differential host tropism, and altered biochemical interaction with cell surface machinery. For example, the Spike mutation S:D614G, found in nearly all Variants of Concern, is associated with higher SARS-CoV-2 loads (MacLean et al., 2020; Yurkovetskiy et al., 2020). Other mutations such as S:N439R, S:N501Y, and S:E484K, have been linked, respectively, to increased transmissibility (Deng et al., 2021), enhanced binding to ACE2 (Starr et al., 2020), and antibody escape (Choi et al., 2020; Greaney et al., 2021). For the vast majority of observed mutations, however, links to SARS-CoV-2 fitness are unknown and functional consequences remain uncharacterized.

Fortunately, the SARS-CoV-2 pandemic has prompted a global genomic surveillance program of unprecedented scope and scale, with more than 10 million virus genomes sequenced to date. This growing quantity of genomic surveillance data provides a unique opportunity to interrogate the dynamics of viral infection and quantify selective forces acting on lineages and mutations. Current methods to analyze such data typically rely on phylogenetic analysis or parametric growth models. Phylogenetic methods usually rely on expensive Markov Chain Monte Carlo (MCMC) for inference, with the result that handling more than $\sim$5000 samples becomes computationally infeasible (Pybus and Rambaut, 2009; Morel et al., 2021). By contrast, parametric growth models are scalable to large datasets but typically do not systematically account for competition between multiple lineages, have little to say about newly emergent lineages, and cannot pinpoint the genetic determinants of differential fitness (Davies et al., 2021; Volz et al., 2021).

Two recently developed methods address some of these shortcomings. The first method, PyR_0, is a hierarchical Bayesian parametric growth model that jointly estimates growth rates for multiple lineages across multiple geographic regions (Obermeyer et al., 2022). Since PyR_0 regresses growth rates against genotype, it can also make inferences about the genetic determinants of differential fitness. Moreover, since PyR_0 relies on variational inference it can be applied to large datasets. However, variational inference also results in poor uncertainty estimates and the parametric likelihood that underlies PyR_0 makes ad hoc assumptions about the noise characteristics of surveillance data. The second method, which we refer to as MAP, likewise regresses growth rates against genotype but instead utilizes an elegant diffusion-based likelihood that is better suited to the stochastic dynamics of viral transmission (Lee et al., 2022). However, unlike PyR_0 MAP does not assume that most alleles are approximately neutral, with the result that MAP risks inferring non-negligible selection effects for implausibly many alleles.

In the following we set out to formulate a method—Bayesian Viral Allele Selection—that combines and improves upon the respective strengths of both PyR_0 and MAP and achieves the following desiderata. First, we retain both methods' scalability

to large datasets and their ability to account for competition between multiple co-circulating lineages. Second, we retain both methods' ability to infer the genetic determinants of differential fitness, which is important both for understanding the biology of transmission and pathogenesis and for predicting the fitness of emergent lineages. Third, we incorporate the sparsity assumption of PyR_0—namely that most alleles are approximately neutral—while adopting the principled diffusion-based likelihood that underlies MAP. Finally we discard variational inference in favor of efficient MCMC so as to obtain more plausible uncertainty estimates, which provide crucial nuance for public health agencies.

To establish the operational characteristics of our method we perform a large suite of simulations, including detailed comparisons to PyR_0, MAP, and another diffusion-based method we introduce (Laplace). We find that Bayesian Viral Allele Selection (BVAS) performs well across the board, with notable advantages of BVAS being its robustness to hyperparameter choices, its satisfactory uncertainty estimates and the fact that it offers interpretable Posterior Inclusion Probabilities that can be used to prioritize alleles for follow-up study.

We apply BVAS to 6.9 million SARS-CoV-2 genomes obtained through April 18th, 2022, noting that, to the best of our knowledge, this is the largest such analysis to date. Our genome wide analysis identifies known functional hot spots in the SARS-CoV-2 genome like the receptor-binding domain (RBD) in the S gene as well as additional hits in regions of the genome whose function is less well understood like the ORF1ab polyprotein. We argue, based on a retrospective backtesting analysis, that running BVAS periodically as part of a real-time genomic surveillance program could provide valuable estimates of the growth rates of new lineages as they emerge. In addition, we conduct an analysis that allows for vaccination-dependent selection effects and find tantalizing evidence that S:N501Y exhibits vaccination-dependent differential fitness. Finally, we conduct an analysis that aims to identify pairs of mutations whose fitness effect is not additive (i.e. epistasis), which likewise points to the important role played by the RBD residue N501.

2. Models and Methods

2.1. Viral Infection as Diffusion

The starting point for both MAP and BVAS is a branching process that encodes the dynamics of infected individuals at time t stochastically generating secondary infections at time $t+1$.[1] Since SARS-CoV and SARS-CoV-2 are known to exhibit super-spreading (Lloyd-Smith et al., 2005; Althouse et al., 2020)—i.e. a minority of infected individuals causes the majority of secondary infections—the number of secondary infections is assumed to be governed by a Negative Binomial distribution, which has a large variance for small values of the dispersion parameter k. In particular we assume that if a given individual is infected with a variant v with reproduction number R_v, the number of secondary infections due to that individual has mean R_v and variance $R_v + R_v^2/k$. If we let $n_v(t)$ denote the total number of individuals at time t infected with variant v, our assumptions result in the following discrete time process:

$$n_v(t+1) \sim \mathrm{NegBin}(\mathrm{mean} = n_v(t) R_v, \mathrm{dispersion} = n_v(t) k) \quad (1)$$

[1] A more detailed exposition of this and the following sections can be found in the supplement.

To connect these dynamics to genotype, we assume that variants are characterized by A alleles and that each variant v is encoded as a binary vector $\mathbf{g}_v \in \{0,1\}^A$. We then express R_v as $R_v = R_0(1+\Delta R_v)$, where R_0 corresponds to the wild-type variant, and assume that ΔR_v is governed by a linear[2] additive model

$$\Delta R_v = \sum_{a=1}^{A} g_{v,a} \beta_a \quad (2)$$

where $\boldsymbol{\beta} \in \mathbb{R}^A$ are allele-level selection coefficients. If we transform from case counts $n_v(t)$ to allele frequencies $x_a(t)$, Lee et al. (2022) show that the dynamics in Eqn. 1 are equivalent to the following diffusion process in allele frequency space

$$\mathbf{x}(t+1) \sim \mathcal{N}(\mathbf{x}(t) + \mathbf{d}(t), \nu^{-1}\boldsymbol{\Lambda}(t)) \quad (3)$$

where $\mathbf{d}(t) \in \mathbb{R}^A$ is the A-dimensional drift, given by

$$d_a(t) = x_a(t)(1-x_a(t))\beta_a + \sum_{b \neq a} \left(x_{ab}(t) - x_a(t)x_b(t)\right)\beta_b \quad (4)$$

The $A \times A$ diffusion matrix $\boldsymbol{\Lambda}(t)$ is given

$$\Lambda_{ab}(t) = x_{ab}(t) - x_a(t)x_b(t) \quad (5)$$

where $x_{ab}(t)$ is the fraction of infected individuals at time t who carry alleles a and b. Finally ν is the effective population size given by

$$\nu \equiv \left(\tfrac{1}{R_0} + \tfrac{1}{k}\right)^{-1} n = \tfrac{kR_0}{k+R_0} n \quad (6)$$

where n is the total number of infected individuals. Importantly, the equivalence of Eqn. 1 and Eqn. 3 holds in the diffusion limit of large n.[3]

2.2. MAP

The simplest model that utilizes the diffusion-based likelihood in Eqn. 3 is formulated as follows (we refer the reader to Lee et al. (2022) for additional discussion). First we place a Multivariate-Normal prior on the selection coefficients $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}|\tau) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \tau^{-1}\mathbb{1}_A) \quad (7)$$

where $\tau > 0$ is the prior precision and $\mathbb{1}_A$ is the $A \times A$ identity matrix. For observed incremental allele frequency changes

$$\mathbf{y}(t) \equiv \mathbf{x}(t+1) - \mathbf{x}(t) \quad (8)$$

the likelihood is given by

$$p(\mathbf{y}_{1:T-1}|\boldsymbol{\beta}, \nu) = \prod_{t=1}^{T-1} \mathcal{N}(\mathbf{d}(t|\boldsymbol{\beta}), \nu^{-1}\boldsymbol{\Lambda}(t)) \quad (9)$$

where we have assumed that ν is constant across time. Since $\boldsymbol{\beta}$ appears linearly in the drift $\mathbf{d}(t|\boldsymbol{\beta})$ and the prior is Multivariate-Normal, the corresponding maximum a posteriori (MAP) estimate is available in closed form:

$$\boldsymbol{\beta}^{\mathrm{MAP}} = \left(\sum_{t=1}^{T-1} \boldsymbol{\Lambda}(t) + \frac{\tau}{\nu}\mathbb{1}_A\right)^{-1} \left(\mathbf{x}(T) - \mathbf{x}(1)\right) \quad (10)$$

An attractive property of this estimator is that it can be computed in $\mathcal{O}(A^3)$ time and is thus quite fast on modern

[2] We consider quadratic effects in Sec. 4.7.

[3] The use of diffusion processes similar to that in Eqn. 3 has a long history in population genetics, including seminal work by Kimura (Kimura, 1964) as well recent applications that employ diffusion-based likelihoods in the context of statistical inference (Lacerda and Seoighe, 2014; Terhorst et al., 2015; Ferrer-Admetlla et al., 2016; Sohail et al., 2021).

hardware, at least for A up to $A \sim 10^4 - 10^5$. An unattractive property of this estimator is that it can perform poorly in the high-dimensional regime, $A \gg 1$, since we expect most alleles to be neutral, but β_a^{MAP} will generally be non-zero for all a.

2.3. Bayesian Viral Allele Selection

We now introduce our method: Bayesian Viral Allele Selection (BVAS). We expect most alleles to be nearly neutral ($\beta_a \approx 0$) and we would like to explicitly include this assumption in our model. To do so we utilize the modeling motif of Bayesian Variable Selection (Chipman et al., 2001):

$$\begin{aligned} &\text{[inclusion variables]} && \gamma_a \sim \text{Bernoulli}(h) \\ &\text{[selection coefficients]} && \boldsymbol{\beta}_\gamma \sim \mathcal{N}(0, \tau^{-1}\mathbb{1}_{|\gamma|}) \\ &\text{[allele frequency changes]} && \mathbf{y}(t) \sim \mathcal{N}(\mathbf{d}(t|\boldsymbol{\beta}_\gamma), \nu^{-1}\boldsymbol{\Lambda}(t)) \end{aligned} \tag{11}$$

where $a = 1, ..., A$ and $t = 1, ..., T-1$. Here each Bernoulli latent variable $\gamma_a \in \{0, 1\}$ controls whether the ath coefficient β_a is included ($\gamma_a = 1$) or excluded ($\gamma_a = 0$) from the model; in other words it controls whether the ath allele is neutral or not. The hyperparameter $h \in (0, 1)$ controls the overall level of sparsity; in particular $S \equiv hA$ is the expected number of non-neutral alleles a priori. The $|\gamma|$ coefficients $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{|\gamma|}$ are governed by a Normal prior with precision τ where $\tau > 0$ is a fixed hyperparameter. Here $|\gamma| \in \{0, 1, ..., A\}$ denotes the total number of non-neutral alleles in a given model.[4]

In addition to inducing sparsity, an attractive feature of the model in Eqn. 11 is that—because it is formulated as a model selection problem—it explicitly reasons about whether each allele is neutral or not. In particular this model allows us to compute the *Posterior Inclusion Probability* or PIP, an interpretable score that satisfies $0 \leqslant \text{PIP} \leqslant 1$. The PIP is defined as $\mathrm{PIP(a)} \equiv \mathrm{p}(\gamma_\mathrm{a} = 1|\mathbf{y}_{1:\mathrm{T}-1})$, i.e. $\mathrm{PIP}(a)$ is the posterior probability that allele a is included in the model. This quantity should be contrasted to h in Eqn. 11, which is the *a priori* inclusion probability. Alleles that have large PIPs are good candidates for being causally linked to viral fitness.

In Eqn. 11 we assume that h is known. An alternative is to place a prior on h, $h \sim \text{Beta}(\alpha_h, \beta_h)$, and infer h from data. See Sec. S9 for details.

2.4. MCMC Inference

BVAS admits efficient MCMC inference via a recently introduced algorithm dubbed Tempered Gibbs Sampling (Zanella and Roberts, 2019). This is quite remarkable: the underlying inference problem is very challenging, since i) it is a transdimensional inference problem defined on a mixed discrete/continuous latent space; and ii) the size of the model space, namely 2^A, is astronomically large. The feasibility of MCMC inference in this setting is enabled by the specific Gaussian form of the diffusion-based likelihood in Eqn. 9 and would be impractical for most other (non-conjugate) likelihoods. Thus BVAS is made possible by a pleasant synergy between the form of the prior and the likelihood.

As we explain in more detail in Sec. S4 the resulting inference algorithm has $\mathcal{O}(|\gamma|^2 A)$ computational cost per MCMC iteration and is thus quite fast on modern hardware. Here $|\gamma|$ is the total number of non-neutral alleles, which by assumption satisfies $|\gamma| \ll A$. Notably the computational complexity does not include terms that are quadratic or cubic in A, since the (strict) sparsity of Bayesian variable selection implies that the required linear algebra never involves $A \times A$ matrices. Importantly, the viability of MCMC inference means that we expect to achieve satisfactory uncertainty estimates, in particular ones that explicitly weigh differing hypotheses about which alleles are neutral and which are not. Indeed the BVAS posterior mean of $\boldsymbol{\beta}$ can be viewed as an evidence-weighted linear combination of 2^A MAP estimates.

[4] In the following we drop the γ subscript on $\boldsymbol{\beta}_\gamma$ to simplify the notation.

2.5. Multiple spatial regions

In the above we have assumed a single spatial region. To apply either BVAS or MAP to multiple spatial regions we simply add a subscript where necessary and form a product of diffusion-based likelihoods for N_R regions indexed by r:

$$\prod_{r=1}^{N_R} \prod_{t=1}^{T-1} \mathcal{N}(\mathbf{d}_r(t|\boldsymbol{\beta}), \nu_r^{-1}\boldsymbol{\Lambda}_r(t)) \tag{12}$$

As discussed in Sec. S4, including multiple regions has negligible impact on the computational cost, since all summations over the region index r are performed once in pre-processing.

2.6. Estimating the effective population size

The likelihood in Eqn. 9 depends on the effective population size ν, a quantity that we do not know a priori and need to estimate from data. For a given region r Eqn. 9 implies

$$\mathbb{E}\left[\mathbf{y}_r(t)^{\mathrm{T}}\mathbf{y}_r(t)\right] = \mathbf{d}_r(t)^{\mathrm{T}}\mathbf{d}_r(t) + \nu_r^{-1}\mathrm{Tr}\boldsymbol{\Lambda}_\mathrm{r}(\mathrm{t}) \tag{13}$$

so that if we assume that the drift term is subdominant we obtain the approximation

$$\hat{\nu}_r \approx \frac{\mathrm{Tr}\boldsymbol{\Lambda}_\mathrm{r}(\mathrm{t})}{\mathbb{E}\left[\mathbf{y}_r(t)^{\mathrm{T}}\mathbf{y}_r(t)\right]} \tag{14}$$

We note that, since $\mathbf{d}_r(t)^{\mathrm{T}}\mathbf{d}_r(t) \geqslant 0$, we would expect $\hat{\nu}_r$ to be an underestimate of ν_r, especially if the effective population size is large. This results in the following simple estimator

$$\hat{\nu}_r = \tfrac{1}{T-1}\sum_{t=1}^{T-1} \frac{\mathrm{Tr}\boldsymbol{\Lambda}_\mathrm{r}(\mathrm{t})}{\mathbf{y}_r(t)^{\mathrm{T}}\mathbf{y}_r(t)} \tag{15}$$

where we have averaged Eqn. 14 over $T-1$ time steps.

To accommodate multiple regions we compute $\hat{\nu}_r$ within each region using Eqn. 15 and then compute a single global effective population size $\hat{\nu}$ by computing the median of $\{\hat{\nu}_r\}$. With this choice all regions contribute equally to the likelihood. See Sec. S6 for additional details and discussion.

2.7. Sampling Rate

As we show in Sec. S8 an attractive property of the diffusion process in Eqn. 3 is that it behaves sensibly in the presence of sampling, i.e. the fact that not all viral sequences are observed in real world datasets. Indeed if sampling is i.i.d. and the sampling rate is ρ with $0 < \rho \ll 1$ then the effect of sampling is to renormalize the effective population size in Eqn. 6 as

$$\nu \to \left(\tfrac{1}{R_0} + \tfrac{1}{k} + \tfrac{2}{\rho}\right)^{-1} n \tag{16}$$

This means that the covariance structure in Eqn. 3 remains intact, which is important because it is precisely this 2nd order information that helps BVAS and MAP disentangle driver mutations from passenger mutations. This is reassuring because for SARS-CoV-2, where even the most ambitious surveillance programs satisfy $\rho \ll k$, the effective population size is dominated by the effects of sampling and $\nu \approx \frac{\rho}{2}n$.

2.8. Vaccination-dependent effects

Suppose we know the vaccination rate $0 \leqslant \phi_r(t) \leqslant 1$ for a given region r. We would like to incorporate this information into our modeling by allowing for vaccination-dependent selection. To do so we write the drift in region r as

$$d_{r,a}(t) = x_{r,a}(t)(1 - x_{r,a}(t))\,(\beta_a + \phi_r(t)\alpha_a) + \sum_{b \neq a} (x_{r,ab}(t) - x_{r,a}(t)x_{r,b}(t))\,(\beta_b + \phi_r(t)\alpha_b) \tag{17}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^A$ is a second group of selection coefficients whose strength is modulated by the time- and region-local vaccination rate. In particular $\boldsymbol{\alpha}$ only has a non-negligible effect on infection dynamics when $\phi_r(t)$ is itself non-negligible. Disentangling the effects of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is difficult a priori. Our hope, however, is that a Bayesian variable selection approach with robust MCMC inference should be up to the task provided we have enough data. See Sec. S7 for additional discussion.

2.9. Alternative Model: Laplace

Finally we describe the simplest modification of MAP that can account for the expected sparsity of non-neutral alleles.[5] In this approach we place a Laplace prior on $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}|\sigma^{\text{Laplace}}) = \tfrac{1}{2\sigma^{\text{Laplace}}} \exp\left(-\frac{||\boldsymbol{\beta}||_1}{\sigma^{\text{Laplace}}}\right) \tag{18}$$

where $||\boldsymbol{\beta}||_1$ is the L^1 norm of $\boldsymbol{\beta}$ and $\sigma^{\text{Laplace}} > 0$ is a hyperparameter that controls the expected level of sparsity. We then define the maximum a posteriori estimate under this Laplace prior:[6]

$$\boldsymbol{\beta}^{\text{Laplace}} \equiv \arg\max_{\boldsymbol{\beta}} p(\boldsymbol{\beta}|\sigma^{\text{Laplace}}) p(\mathbf{y}_{1:T-1}|\boldsymbol{\beta}, \nu) \tag{19}$$

This estimator cannot be computed in closed form but can be readily approximated with iterative optimization techniques. We will consider Laplace alongside BVAS, MAP, and PyR_0 in our simulations, which we turn to next.

3. Simulation Results

To assess the performance of our method we conduct an extensive suite of simulation-based experiments, including experiments that rely solely on simulated data as well as a semi-synthetic experiment that relies on perturbed SARS-CoV-2 data.

3.1. Simulation details

Our simulator closely follows the structure of the discrete time process in Eqn. 1. The most salient details are as follows (see Sec. S13 for details). We include exactly 10 non-neutral alleles of varying effect size, with typical reproduction numbers for variants v ranging between 0.9 and 1.1. In each simulation we consider a given number of N_R regions and $T = 26$ time steps. The initial number of infected individuals at time $t = 1$ within each region is drawn from a Negative Binomial distribution with mean 10^4. Case counts for $t = 2, ..., T$ are determined by the stochastic dynamics in Eqn. 1 with $k = 0.1$. This value of k is chosen since it is consistent with estimates of the SARS-CoV-2 dispersion parameter (Lau et al., 2020; Endo et al., 2020; Bi et al., 2020; Miller et al., 2020). These raw counts are then subjected to Binomial sampling with mean $\rho = 0.01$, i.e. the viral sequences of 99% of cases are not observed. Thus our parameter choices result in simulated data that are highly stochastic and that constitute a regime in which we expect that recovering the true selection coefficients $\boldsymbol{\beta}^*$ is quite challenging. Unless noted otherwise, we generate 20 datasets per condition. We make these choices because they result in simulated data that exhibit some of the characteristics of our SARS-CoV-2 data. In particular, typical estimated effective population sizes $\hat{\nu}$ range from about 25 to about 140 with a mean of about 75.

3.2. Method Comparison

We compare four methods for inferring allele-level selection using simulated data, in particular three diffusion-based methods (MAP, BVAS, and Laplace) and PyR_0. For all methods except for BVAS we rank allele-level hits by the absolute effect size, whereas for BVAS we rank by the Posterior Inclusion Probability (PIP). See Sec S13.3 for the hyperparameter choices made.

In Figure 1 we report results on the hit rate, which we define as the fraction of the top 10 hits that are causal.[7] This metric is convenient since it does not depend on any method-specific threshold for calling hits. As expected the hit rate generally increases as the number of regions increases and decreases as the number of alleles increases (since the number of possible spurious hits increases). Strikingly, BVAS exhibits the best hit rates across the board. Laplace and MAP are competitive with BVAS in some regimes, but their performance degrades in other regimes, particularly when the number of alleles is large.

We hypothesize that the main reason for the poor performance of MAP in some regimes is the fact that MAP does not enforce sparsity in the allele-level coefficients $\boldsymbol{\beta}$. This effect is particularly evident from the mean absolute error (MAE) results in Figure 2, where it can be seen that the MAP MAE is large across the board, since MAP assigns non-negligible effect sizes to a large number of alleles. As the number of regions and thus the total amount of data increases, MAP tends to identify ever more non-negligible effects, potentially leading to a large number of spurious hits.

In contrast to MAP, PyR_0 and Laplace do impose sparsity on the allele-level coefficients $\boldsymbol{\beta}$. We hypothesize that one of the main reasons for the poor performance of PyR_0 and Laplace in some regimes is the fact that they rely on hyperparameters that are difficult to choose. This is especially the case for PyR_0, which contains 7 model hyperparameters, the most important of which is a direct analog to σ^{Laplace}.

To make this broader point concrete we investigate the sensitivity to the Laplace regularization scale σ^{Laplace} in Figure 3. We find that moderate changes in σ^{Laplace} lead to significant degradation in performance. Since there is no principled method to choose σ^{Laplace} a priori, one must instead rely on simulation-based intuition. Since, however, any simulation cannot capture all the effects that characterize real data and since it is unclear a priori what simulation parameters should be used, it remains difficult to choose σ^{Laplace} and so the

[5] For additional discussion please refer to Sec. S10, where we describe several alternative models that make use of the diffusion-based likelihood in Eqn. 3.

[6] For the sake of precision we should probably refer to MAP as MAP-Gaussian and the approach described here as MAP-Laplace. However, for brevity we instead refer to these methods as MAP and Laplace, respectively.

[7] Causal alleles are those for which the true effect is non-zero.

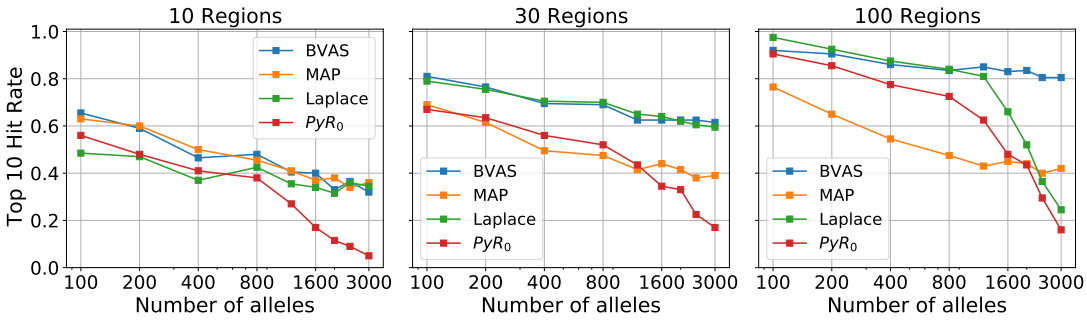

Fig. 1. We compare the hit rate for four different methods using simulated data, where the hit rate is defined as the fraction of the top 10 hits that are causal. Results are averaged across 20 independent simulations. See Sec. 3.2 for discussion.

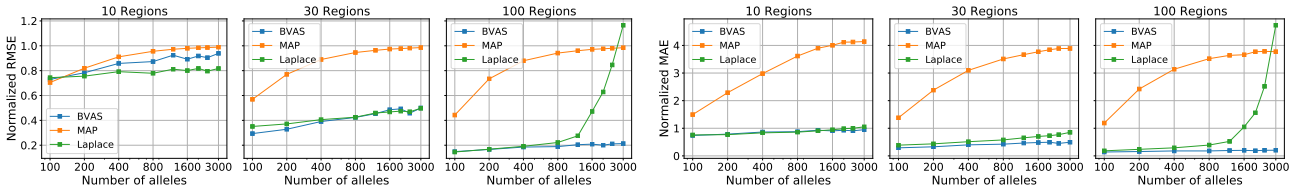

Fig. 2. We report the accuracy of inferred selection coefficients $\boldsymbol{\beta}$ for three diffusion-based methods using the simulated data described in Sec. 3.1. We consider two metrics: root mean squared error (RMSE; left) and mean absolute error (MAE; right). In both cases the metric is normalized such that the value of the metric for $\boldsymbol{\beta} = \mathbf{0}$ is equal to unity. For example, the RMSE is normalized by $||\boldsymbol{\beta}^*||_2$, where $\boldsymbol{\beta}^*$ are the true effects. We do not include a comparison to PyR_0, since it utilizes a somewhat different likelihood, making direct comparison subtle. See Sec. 3.2 for discussion.

sensitivity in Figure 3 is troubling. In the next section we show that BVAS exhibits less sensivity to hyperparameter choices.

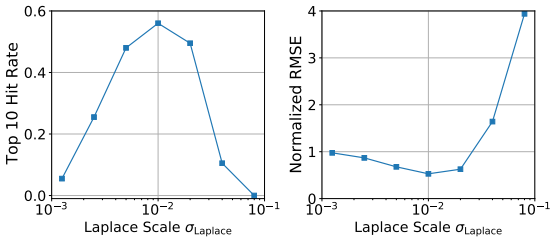

Fig. 3. We explore the sensitivity of the Laplace method to the prior scale σ^{Laplace}. Changing σ^{Laplace} from the optimal value of $\sigma^{\text{Laplace}} \approx 0.01$ results in significantly worse performance. We consider $A = 3000$ alleles and $N_R = 30$ regions and generate 40 simulated datasets.

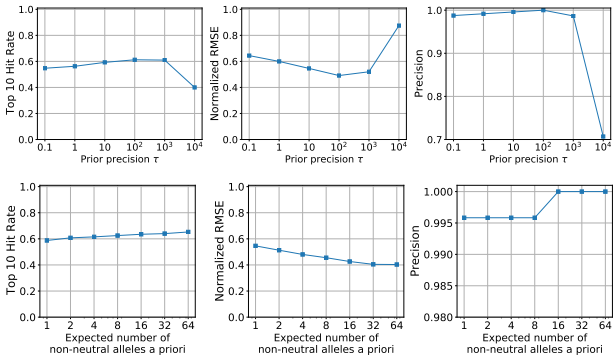

Fig. 4. We explore the extent to which BVAS performance is sensitive to its two hyperparameters, namely τ and $S = hA$, using the simulated data described in Sec. 3.1. We simulate data for $N_R = 30$ regions and $A = 3000$ alleles and generate 40 datasets. See Sec. 3.3 for discussion.

3.3. BVAS sensitivity to hyperparameters and $\hat{\nu}$

BVAS is specified by two hyperparameters: the prior inclusion probability h and the prior precision τ.[8] The quantity $\tau^{-\frac{1}{2}}$ controls the expected scale of effect sizes β. For example, for $\tau = 100$ the prior standard deviation of β is 0.1. This choice implies that $\sim 95\%$ of prior probability mass concentrates on the range $\beta \in [-0.2, 0.2]$. In Figure 4 (top row) we depict the sensitivity of BVAS to changes in τ. We find that the sensitivity to τ is small over about 4 orders of magnitude. It is only for very large τ ($\tau = 10^4$) that we see a large drop in performance.

In Figure 4 (bottom row) we depict the sensitivity of BVAS to changes in $S \equiv hA$, which is the expected number of non-neutral alleles a priori. We find only moderate sensitivity as S ranges from $S = 1$ to $S = 64$. In other words it is not necessary for S to be an accurate estimate of the number of non-neutral alleles (10 in our simulations): the posterior is a compromise between the prior and the likelihood and for reasonable choices of S the likelihood can overwhelm the prior if there is sufficient evidence for non-neutral alleles. Importantly the precision remains high for all values of S. The effect of choosing small S is to be more conservative; in particular some weak effects at the threshold of discovery may be assigned small PIPs. This robustness to changes in S is reassuring because our a priori knowledge of the number of non-neutral alleles in real data is limited.

[8] For an extended dicussion of h see Sec. S9. Note that if a prior is placed on h the hyperparameters instead become $\{\tau, \alpha_h, \beta_h\}$.

Next we explore the sensitivity of BVAS to accurate estimation of the effective population size ν. Note that unlike S or τ, which appear in the prior in Eqn. 11, ν appears in the likelihood. The value of ν evidently plays an important role because it controls the level of noise in the diffusion process. Large values of ν imply that allele frequency increments $\mathbf{y}(t)$ are largely determined by (deterministic) drift. Conversely, small values of ν imply that $\mathbf{y}(t)$ exhibits significant (stochastic) variability that dominates the drift. Thus, with all else equal, increasing ν places more emphasis on fitting the observed apparent drift with the result that BVAS will tend to identify more signal, i.e. more alleles with non-negligible PIPs. Conversely, decreasing ν places less emphasis on fitting the observed apparent drift with the result that BVAS will tend to identify less signal, i.e. fewer alleles with non-negligible PIPs.

We investigate this effect quantitatively in Figure 5, which confirms our intuition. At least for our simulated data the consequences of underestimating ν are more severe than the consequences of overestimating ν; for example if we underestimate ν by a factor of 4 the hit rate drops by a factor of one half. By contrast overestimating ν by a factor of 4 actually improves the hit rate in this simulation, since the tighter likelihood encourages BVAS to seek out less sparse solutions, which results in additional hits for alleles at the margin of discovery. Overall the behavior in Figure 5 is encouraging, since we can estimate the effective population size with moderate accuracy in simulation (see Sec. S13.2). In practice of course we expect worse performance in the context of real data because the noise structure of real data will not precisely follow the noise structure assumed by our diffusion-based likelihood. Nevertheless the fact that the results in Figure 5 exhibit a good degree of robustness for ν estimates that are off by a factor of ~ 2 suggests that running BVAS on real data should be relatively robust to the $\hat{\nu}$ estimation strategy used.

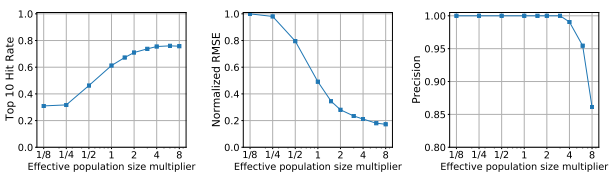

Fig. 5. We explore the extent to which BVAS performance is sensitive to accurate estimation of the effective population size ν using the simulated data described in Sec. 3.1. To do so we modulate our estimate for ν by the indicated multiplier, e.g. $\hat{\nu} \to 2\hat{\nu}$. We simulate data for $N_R = 30$ regions and $A = 3000$ alleles and generate 40 datasets. See Sec. 3.3 for discussion.

3.4. The value of PIPs

In contrast to the other methods we consider BVAS provides a Posterior Inclusion Probability for each allele. In Figure S3 we demonstrate the value of PIPs by exploring the allele-level precision and sensitivity that are obtained if we declare alleles with a PIP above a threshold of 0.1 as hits. We observe very high precision across the board. In other words, if an allele has a high PIP there is good reason to believe it is causally linked to viral fitness, at least if we believe the generative process that underlies our diffusion-based likelihood. It is worth emphasizing that an allele with a moderate effect size can still exhibit a large PIP, thus signifying strong evidence for being causal.

3.5. Variability due to sampling rate

As discussed in Sec. 2.7 our diffusion-based likelihood, Eqn. 9, naturally accommodates random sampling where only a fraction ρ of infected individuals have their viral genomes sequenced. To explore the effects of sampling we generate data with a sampling rate that ranges between 1% and 64%. We find that the results are remarkably robust (see Figure S4), even as the effective population size decreases by a factor of ~ 15 as ρ decreases from 64% to 1% (see Eqn. 16).

3.6. Including vaccination-dependent effects

We now incorporate vaccination rates $\phi_r(t)$ into our simulations, assuming that $\phi_r(t)$ starts at zero everywhere and increases linearly over time. We assume 20 non-zero effects, half of which are vaccination-dependent. Otherwise our simulation follows the specifications of Sec. 3.1. See Figure S5 for results. As we would expect, robustly identifying causal mutations is harder in this setting, and the hit rate for vaccination-dependent effects is lower than for all effects. Nevertheless the precision is high in all cases, which gives us confidence that high PIP vaccination-dependent alleles identified in real data may be causally linked to vaccination-dependent differential fitness.

3.7. Spike-in experiment

We conduct a semi-synthetic experiment where we add 200 spurious alleles to 3000 SARS-CoV-2 lineages (we use data from January 20th 2022 for a total of $A = 2904 + 200 = 3104$ alleles). Each lineage is assigned a Binomial number of non-wild-type spiked-in alleles with mean 2. Since these assignments are independent and identically distributed, the spiked-in alleles are not correlated with the pre-existing genotype in any way and thus any apparent selection effects due to these alleles are due to chance alone. See Figure 6 for results. We find that PyR$_0$ and BVAS select the fewest number of spiked-in alleles, whereas MAP and Laplace select the most. Note that we expect some small number of spiked-in alleles to be selected due to random chance alone. Importantly, across 30 replications none of the methods identifies a single spiked-in allele in the top 20 scoring hits. This is encouraging, since it suggests that the top scoring hits from all four methods should be enriched with causal alleles.

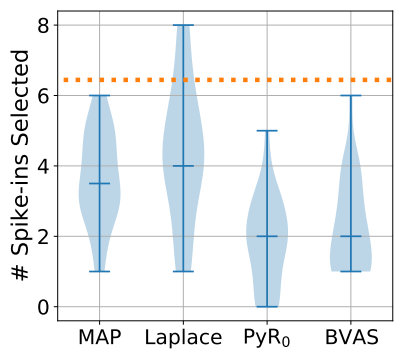

Fig. 6. We compare the robustness of four methods for inferring allele-level selection effects to the addition of spiked-in alleles. We depict the total number of spiked-in alleles that are among the top 100 scoring alleles, where the horizontal line within each violin plot denotes the median and for consistency we rank alleles by the absolute value of the selection coefficient β. The orange dotted line corresponds to the number of spiked-in alleles that would be expected among the top 100 alleles if alleles were ranked at random. We report results from 30 independent simulations.

4. SARS-CoV-2 Analysis

4.1. Data

Our raw data consist of 8.6 million samples downloaded from GISAID (Elbe and Buckland-Merrett, 2017) on April 18th, 2022. In initial pre-processing we follow the procedure in Obermeyer et al. (2022), which relies on a phylogenetic tree constructed by UShER (Turakhia et al., 2021; McBroome et al., 2021), and results in $L = 3000$ SARS-CoV-2 clusters that are finer than the 1662 PANGO lineages in the data (Rambaut et al., 2020). In our main analysis we consider $A = 2975$ non-synonymous amino acid substitutions, excluding both insertions and deletions due to limitations of UShER, and taking Wuhan A as the reference genotype, i.e. $R_0 \equiv R_A$. After filtering to well-sampled regions there remain 6.9 million samples from $N_R = 128$ regions. Allele frequencies for each region are computed in time bins of 14 days and the effective population size is estimated using the global strategy described in Sec. 2.6. Vaccination data for the analysis in Sec. 4.6 are obtained from OWID (Ritchie et al., 2020). For additional details on data pre-processing see Sec. S14.1.

4.2. Fitness of SARS-CoV-2 Lineages and Mutations

We use BVAS to rank the relative fitness of all SARS-CoV-2 lineages. To do so, we fit our model to allele frequencies of 2975 alleles across 128 regions, with $\tau = 100$ and $S = 50$ (so $h = S/A \approx 0.017$) reflecting our prior assumptions that a relatively modest number of non-neutral alleles with (possibly) moderately large selection effects are driving evolution of SARS-CoV-2 fitness.

In Table 1, we report relative growth rate estimates R/R_A for the top 20 lineages. Fitness estimates are broadly concordant with the observed pandemic, with the fittest lineages all Omicron variants. BVAS accurately captures the hierarchy of replacement by fitter lineages with Omicron (BA.2) > Omicron (BA.1) > Delta > Alpha > wild-type virus (Table 2). Notably some PANGO lineages (e.g. B.1.1) exhibit very diverse genotypes and thus correspondingly diverse growth rates, see Figure S7.

Analysis of sublineage fitness reveals that Omicron has fractured into many sublineages whose fitness has increased modestly over time (Table 1). BA.2.12.1 appears to be the fittest lineage observed to date, although many BA.2 sublineages are comparably fit. BA.4 and BA.5, which appear to be descended from BA.2, have recently emerged in South Africa and are reported to have enhanced fitness (Tegally et al., 2022) and additional immune escape (Khan et al., 2022; Cao et al., 2022) relative to Omicron BA.1. Since BVAS regresses growth rate against genotype, it is able to infer that these lineages are among the fittest lineages circulating despite the fact that very few BA.4 and BA.5 sequences are in our dataset. Like BA.2.12.1, BA.4 and BA.5 also possess mutations at Spike position 452 mutations (L452R), underscoring the key role that this site plays in SARS-CoV-2 fitness.

We report the fitness of recombinant lineages in Table S1. In contrast to highly fit lineages that emerged in the BA.2 clade, several recombinants, including those that represent recombination between Delta and BA.1 and BA.1 and BA.2, have been the source of international concern (Colson et al., 2022; Jackson et al., 2021; VanInsberghe et al., 2021). The fittest recombinants are XN and XT, though their fitness is intermediate to that of BA.2 and BA.1. While the appearance of recombinant lineages is striking, the fitness of existing XA - XT recombinants suggests that these particular lineages are unlikely to play an important role in the future.

We report the fitness of top-scoring mutations in Table 3 and plotted along the length of the genome in Figure 7. The strongest signal of selection is in Spike, with the greatest concentration of hits located in the receptor-binding domain (RBD). Strong signals of selection are also observed in the N-terminal domain (NTD) and furin cleavage sites. By effect size, S:L452R is the top-scoring hit. This mutation is found in BA.4/BA.5 and was also an important component of the 'California' variants, B.1.427 and B.1.429. Deng et al. (2021) have shown that this mutation increases infectivity, while Li et al. (2020) and Liu et al. (2021) have shown that it promotes antibody escape. The closely related mutation S:L452Q, one of two key Spike mutations in the fast-growing BA.2.12.1 variant, is also highly ranked, underscoring the importance of this site.

The top-scoring mutations cluster together in various regions of Spike (Figure 9A), particularly the ACE2 binding interface of the RBD (Figure 9B). Three of the top five hits (S:L452R, S:E484K, and S:N501Y) are within the RBD, and a fourth, S:R346K, is just adjacent to it. The mechanisms driving positive selection at these sites likely include increased affinity to ACE2, as was shown for S:N501Y (Starr et al., 2020), as well as escape from neutralizing antibodies that bind in this region, e.g. for S:E484K, S:N440K and S:R346S (Weisblum et al., 2020; Iketani et al., 2022).

We characterized antibody escape of Spike mutations further by correlating BVAS RBD estimates to predictions made by the antibody-escape calculator in Greaney et al. (2022) (Figure S8). This escape calculator is based on deep mutational scanning data for 33 neutralizing antibodies elicited by SARS-CoV-2 and thus represents an independent source of experimental data. As in Obermeyer et al. (2022), there is a strong correlation ($\rho_{\text{Spearman}} = 0.89$) between the two sets of predictions, lending support to BVAS results for the RBD. We assessed the temporal progression of selection effects in SARS-CoV-2 lineages by aggregating ΔR estimates due to S gene, RBD, and non-S-genes contributions (Figure 10). The elevated contribution of S-gene mutations (notably in the RBD) over non-S-gene mutations starting around November 2021 is apparent, in agreement with the results from Obermeyer et al. (2022). Collectively these two results suggest that immune escape has become an increasingly prominent factor in SARS-CoV-2 evolution over time, likely a result of rising rates of convalescent and vaccine-induced immunity to Spike. The correlation between mutations that confer antibody escape and mutations associated with fitness supports the hypothesis that coronaviruses evolve through positive selection of receptor-binding domain mutations that escape neutralizing antibody responses (Kistler and Bedford, 2021).

Other hotspots within Spike include the furin cleavage site, which features three of the top 20 mutations (S:P681R, S:P681H, and S:N679K). These substitutions add positive charge at the cleavage site, likely facilitating S1/S2 cleavage and promoting infectivity by enhancing cell-cell fusion, which has been demonstrated by several groups (Saito et al., 2022; Mohammad et al., 2021; Lista et al., 2021).

BVAS also identifies numerous residues under selection outside of Spike. These include P13L in the N-terminal region and P199L and R203M in the linker region between the N- and C-terminal domains (Figure S9). These latter two mutations have been demonstrated by Syed et al. (2021) to enhance viral packaging. Within ORF1a, several amino acid changes in NSP4 score highly, including T3090I, L3201P, and T3255I (Figure S10). The importance of these effects can be quantified on a per-ORF basis using PIPs, which provide a convenient numerical measure

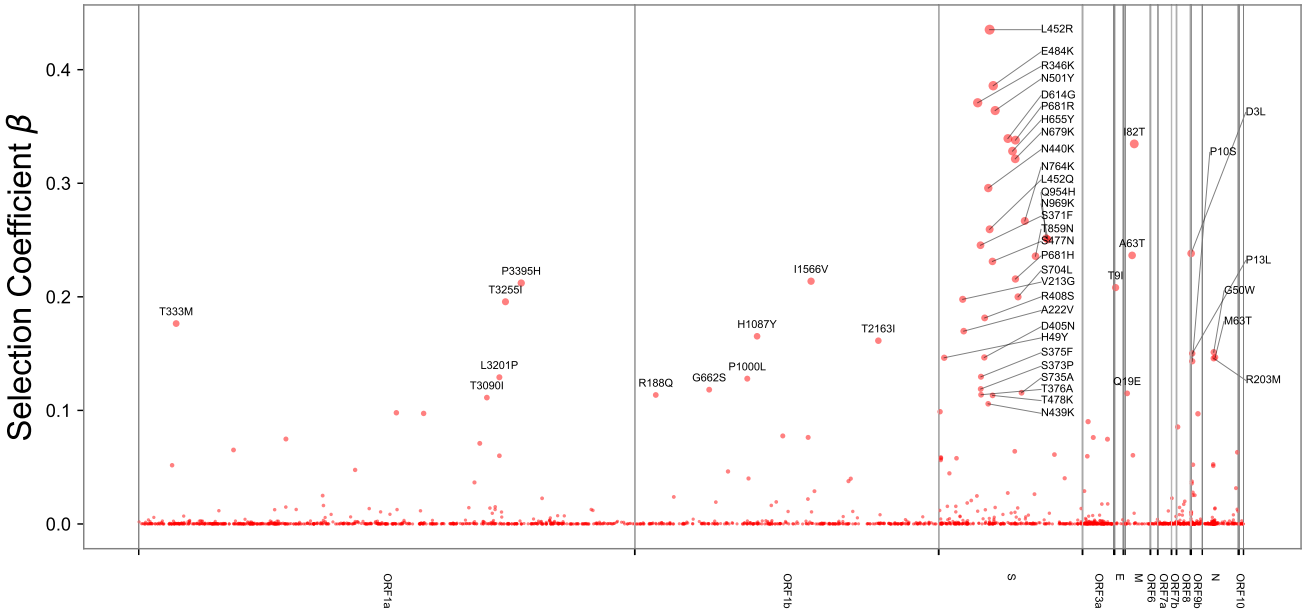

Fig. 7. We depict BVAS hits as a Manhattan plot of the entire SARS-CoV-2 genome, with gene boundaries indicated on the horizontal axis. We do not include the small number of hits with negative selection coefficients (in particular ORF1b:K1383R, ORF1a:A2554V, ORF1b:S959P, and ORF1a:G697R).

of the total evidence for selection due to each allele. To assess the overall contribution of various regions of the SARS-CoV-2 genome to differential viral fitness we sum PIPs across different regions, see Figure S11, quantifying the relative importance of S, N, and several non-structural proteins in ORF1ab.

	Lineage	Growth Rate		Lineage	Growth Rate
1	BA.2.12.1	7.858 ± 0.727	11	BA.2.4	7.399 ± 0.657
2	BA.2.11	7.835 ± 0.667	12	BA.2.7	7.398 ± 0.657
3	BA.5	7.764 ± 0.688	13	BA.2.6	7.398 ± 0.657
4	BA.2.12	7.599 ± 0.703	14	BA.2.5	7.398 ± 0.657
5	BA.2.9.1	7.400 ± 0.659	15	BA.2.1	7.398 ± 0.656
6	BA.2.3.2	7.400 ± 0.657	16	BA.2.13	7.398 ± 0.659
7	BA.2.3.1	7.399 ± 0.657	17	BA.2.2	7.394 ± 0.659
8	BA.2.8	7.399 ± 0.657	18	BA.2	7.370 ± 0.681
9	BA.2.3	7.399 ± 0.656	19	BA.2.15	7.339 ± 0.677
10	BA.2.14	7.399 ± 0.657	20	BA.2.16	7.339 ± 0.676

Table 1. The 20 SARS-CoV-2 lineages with the highest (relative) growth rates R_v/R_A as estimated by BVAS. Here and elsewhere uncertainty estimates are 95% credible intervals.

	WHO Classification	Growth Rate
BA.2	Omicron	7.370 ± 0.681
B.1.1.529	Omicron	6.137 ± 1.259
BA.1	Omicron	5.752 ± 0.683
B.1.617.2	Delta	3.503 ± 0.548
B.1.621	Mu	2.909 ± 0.206
P.1	Gamma	2.495 ± 0.189
B.1.1.7	Alpha	2.262 ± 0.230
B.1.351	Beta	2.135 ± 0.171
B.1.1		1.947 ± 1.544
B.1		1.634 ± 0.844

Table 2. We report the relative growth rates R/R_A of selected SARS-CoV-2 lineages as estimated by BVAS.

4.3. Sensitivity analysis

We perform an extensive sensitivity analysis to better understand the robustness of our results to hyperparameter and data pre-processing choices. In Figure S13-S17 we explore the sensitivity of BVAS growth rate estimates for the fittest lineages, noting that we might expect these estimates to be sensitive to the strength of regularization. We find minimal sensitivity to S (the number of non-neutral alleles expected a priori), the total number of regions included in the analysis N_R, as well as $N^{14}_{\min}$, which controls which time bins enter the analysis, see Figure S13-S15. The sensitivity to the prior precision τ is somewhat larger (see Figure S16), especially as we increase τ to $\tau = 400$, although we would argue that this is an unreasonable prior choice, as it makes even relatively moderate selection effects like $\beta \sim 0.15$ a priori unlikely. Not surprisingly, the sensitivity to the estimated effective population size, $\hat{\nu}$, is fairly large (see Figure S17), roughly comparable to the scale of the underlying statistical uncertainty.

We adopt a more global view in Figure S18-S27, reporting sensitivity analyses for all PIPs and β estimates as well as for growth rates for all 1662 PANGO lineages. Globally, growth rate estimates are remarkably stable with Pearson correlation coefficients of 0.999 or larger. Selection coefficient estimates β are also quite stable, with Pearson correlation coefficients of 0.9 or greater, with the largest sensitivity being again to τ and $\hat{\nu}$. Results for allele-level PIPs are broadly comparable, although they tend to exhibit more variability and outliers, especially

	Mutation	PIP	Beta
1	S:R346K	1.0000	0.371 ± 0.048
2	S:L452R	1.0000	0.435 ± 0.103
3	S:E484K	1.0000	0.386 ± 0.097
4	S:N501Y	1.0000	0.364 ± 0.118
5	M:I82T	1.0000	0.335 ± 0.068
6	S:D614G	1.0000	0.339 ± 0.119
7	S:A222V	1.0000	0.170 ± 0.058
8	S:P681H	1.0000	0.216 ± 0.103
9	S:S477N	0.9999	0.231 ± 0.146
10	S:P681R	0.9999	0.338 ± 0.146
11	ORF1b:K1383R	0.9994	-0.106 ± 0.045
12	S:N440K	0.9992	0.296 ± 0.126
13	ORF1b:H1087Y	0.9978	0.165 ± 0.057
14	ORF1a:A2529V	0.9977	0.097 ± 0.039
15	S:H655Y	0.9939	0.328 ± 0.154
16	ORF1a:A2554V	0.9544	-0.115 ± 0.060
17	S:N679K	0.9500	0.321 ± 0.233
18	ORF1a:T3255I	0.9468	0.196 ± 0.145
19	S:T859N	0.9454	0.236 ± 0.182
20	N:D3L	0.9145	0.238 ± 0.166
21	S:L452Q	0.8827	0.259 ± 0.243
22	S:N764K	0.8334	0.267 ± 0.302
23	M:A63T	0.8127	0.236 ± 0.263
24	S:S371F	0.8065	0.245 ± 0.289
25	S:N969K	0.7888	0.250 ± 0.315

	Mutation	PIP	Beta
1	S:L452R	1.0000	0.435 ± 0.103
2	S:E484K	1.0000	0.386 ± 0.097
3	S:R346K	1.0000	0.371 ± 0.048
4	S:N501Y	1.0000	0.364 ± 0.118
5	S:D614G	1.0000	0.339 ± 0.119
6	S:P681R	0.9999	0.338 ± 0.146
7	M:I82T	1.0000	0.335 ± 0.068
8	S:H655Y	0.9939	0.328 ± 0.154
9	S:N679K	0.9500	0.321 ± 0.233
10	S:N440K	0.9992	0.296 ± 0.126
11	S:N764K	0.8334	0.267 ± 0.302
12	S:L452Q	0.8827	0.259 ± 0.243
13	S:Q954H	0.7888	0.252 ± 0.315
14	S:N969K	0.7888	0.250 ± 0.315
15	S:S371F	0.8065	0.245 ± 0.289
16	N:D3L	0.9145	0.238 ± 0.166
17	M:A63T	0.8127	0.236 ± 0.263
18	S:T859N	0.9454	0.236 ± 0.182
19	S:S477N	0.9999	0.231 ± 0.146
20	S:P681H	1.0000	0.216 ± 0.103
21	ORF1b:I1566V	0.6978	0.214 ± 0.321
22	ORF1a:P3395H	0.6961	0.212 ± 0.321
23	E:T9I	0.6931	0.208 ± 0.287
24	S:S704L	0.7629	0.200 ± 0.257
25	S:V213G	0.6662	0.198 ± 0.316

Table 3. The top 25 fitness-associated SARS-CoV-2 mutations as estimated by BVAS and ranked by PIP (left) and β (right). The two rankings are largely the same, with 19 mutations appearing in both. Uncertainty estimates are 95% credible intervals.

for alleles with smaller PIPs. Importantly concordance between independent MCMC chains is exceptionally high ($R \geqslant 0.9997$), see Figure S18, suggesting that MCMC error is minimal.

4.4. Backtesting

We perform a backtesting analysis in which we run BVAS on subsets of the data defined by varying end dates. Doing so allows us to assess the possible benefits of running BVAS periodically as part of a real-time genomic surveillance program.[9] See Figure S28 for results. We find that by May 13th 2021 BVAS predicts that various Delta sublineages are fitter than B.1.1.7 (Alpha), which was the most prevalent lineage in England and elsewhere at the time. Similarly, we find that by December 8th 2021 BVAS predicts that various Omicron sublineages are fitter than AY.4.2.1 (Delta). Notably, since BVAS regresses R_v against genotype, we also obtain plausible estimates for newly emergent lineages that have only been sequenced a small number of times. Finally by January 12th 2022 BVAS predicts that BA.2 was fitter than BA.1, a prediction that has been borne out by the subsequent takeover of BA.2.

During the time periods considered in our backtesting analysis the number of samples of these newly emergent lineages was increading rapidly, with the result that BVAS growth rate estimates increase markedly as more data become available and it becomes increasingly clear that these lineages were the fittest lineages yet observed. Importantly, estimates stabilize after sufficient data have been collected, see Figure S29.

[9] We caution that these results need to be interpreted with care, since in practice it can take several weeks before individual genomic samples are deposited in GISAID. Additionally, the calling of new lineages is also associated with a time lag.

4.5. Comparison to MAP and PyR$_0$

We perform detailed comparisons of BVAS to both MAP (Lee et al., 2022) and PyR$_0$ (Obermeyer et al., 2022). Here we provide a brief summary; see Sec. S14.4-S14.5 and Figures S30-S40 for a detailed discussion. Despite differences in methodology, results from BVAS, MAP and PyR$_0$ are in broad *qualitative* agreement, suggesting that all three methods are capable of identifying (at least some) leading driver mutations in SARS-CoV-2. Because it is only very recently that genomic surveillance data have become available at this scale and the corresponding analysis methods are still in their infancy, we believe this finding is encouraging for this emerging new field.

While it is difficult to definitively establish the superiority of one method over another without a larger corpus of experimental data to serve as ground truth, we believe that the advantages of BVAS become apparent upon closer comparison. First, inferred selection effects are much sparser in the case of BVAS, which aids interpretability and is arguably more plausible a priori. Selection coefficients inferred by MAP are very dense unless the regularization parameter $\gamma_{\text{reg}} = \tau/\nu$ is made sufficiently large. However, in this limit growth rate estimates appear to be over-regularized towards Wuhan A. Second, uncertainty estimates for MAP and (especially) PyR$_0$ are much narrower than for BVAS,

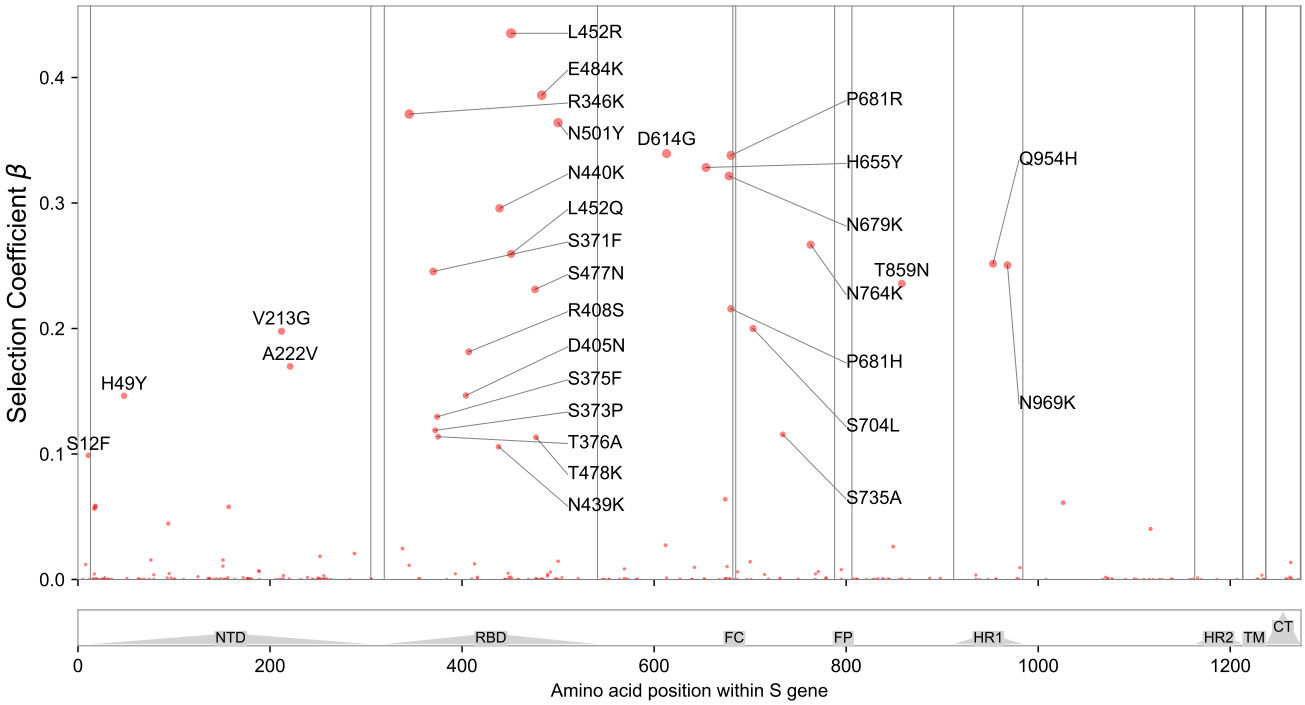

Fig. 8. View of the 1237 amino acids of the S protein, annotated by structure (Huang et al., 2020); many top-scoring mutations are located in the N-terminal domain (NTD), receptor-binding domain (RBD), and furin cleavage (FC) site. Regions containing the fusion peptide (FP), heptad repeat (HR) 1 and 2, transmembrane domain (TM), and C-terminal domain (CTD) are also annotated.

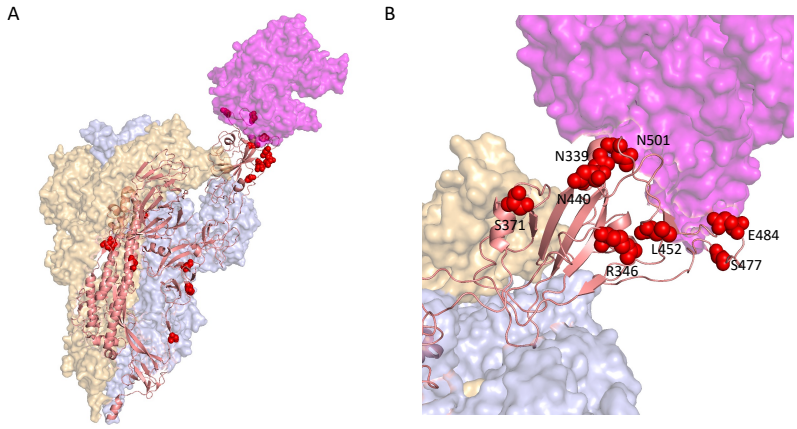

Fig. 9. A. We depict the locations of the top 20 Spike hits, ranked by PIP, on the Cryo-EM structure of a Spike trimer bound to ACE2 (magenta) at 3.9 Angstrom resolution in the single RBD "up" conformation from (Zhou et al., 2020) **B.** Enlarged view of the RBD-ACE2 interface, showing the spatial proximity of S:R346, S:N339, S:N440, S:L452, S:S477, S:E484, and S:N501.

especially for allele-level quantities like selection coefficients. Finally, BVAS exhibits much less sensitivity to hyperparameter choices than PyR_0, which—together with rigorous sparsity requirements—is one of the key factors contributing to the strong performance of BVAS in simulations.

4.6. Vaccination-dependent effects

We incorporate independent data on vaccination rates from 127 regions compiled by OWID.[10] This has little impact on estimates for most vaccination-independent effects (Figure S42), i.e. BVAS finds that selection effects in the data are largely explainable by vaccination-independent effects. Indeed we find only two alleles with large PIPs in the vaccination-dependent model (Table 4).

The strongest evidence for allele-level dependence on vaccination rates is in S:N501Y, which is also the only allele for which BVAS finds it important to include both a vaccination-dependent effect α *and* a vaccination-independent effect β. In our analysis the selection coefficient estimates for S:N501Y are $\alpha \approx -0.15$ and $\beta \approx 0.41$. This means that the model predicts a selection effect due to the S:N501Y allele of $\beta \approx 0.41$ in a

[10] See Sec. S7 and Sec. S14.1.6 for additional details.

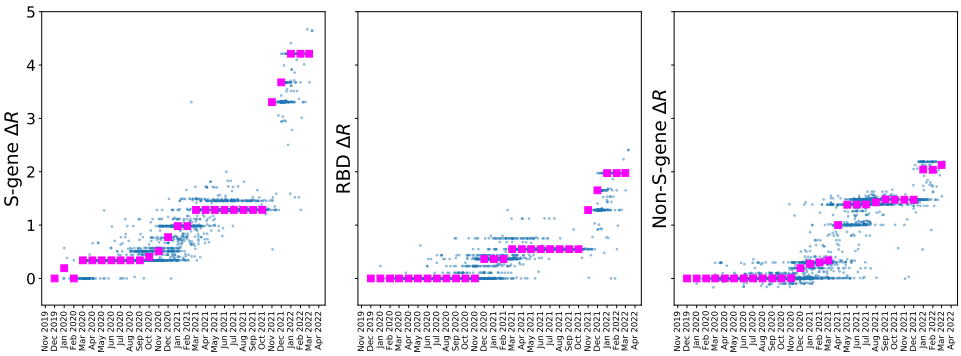

Fig. 10. We aggregate BVAS β estimates into S-gene, RBD (S gene receptor-binding domain), and non-S-gene components for $L = 3000$ SARS-CoV-2 clusters (blue points). The horizontal axis denotes the date at which each cluster first emerged, and magenta squares denote the median ΔR within each monthly bin. The elevated contribution of S-gene mutations (notably in the RBD) over non-S-gene mutations starting around November 2021 is conspicuous. Compare to Figure 2CDE in Obermeyer et al. (2022).

completely unvaccinated population and a selection effect of $\alpha + \beta \approx 0.26$ in a completely vaccinated population. This can be interpreted as saying that vaccination appears to confer differential protection against the S:N501Y allele, i.e. on top of the protection that is conferred against typical SARS-CoV-2 variants that do not carry S:N501Y. Notably, S:N501Y also exhibits a large PIP in an analysis conducted with a different definition of vaccination rate (Table S3).

These results are also supported by a direct analysis of raw allele frequencies of S:N501Y. Indeed S:N501Y exhibited a rapid rise and fall in prevalence in Spring 2021 (Figure S43), at the same time as vaccination rates were ramping up in many of the regions in our dataset. Moreover, while the behavior of S:N501Y is partially explained by the rise and fall of B.1.1.7 (Alpha), S:N501Y came to prominence in some regions (notably Brazil) via P.1 (Gamma) where B.1.1.7 was never dominant. Finally, the change in frequency of S:N501Y is correlated to vaccination status (Figure S44); this correlation is stronger than the correlation between changes in B.1.1.7 frequency and vaccination status.

S:N501Y thus appears to exhibit vaccination-dependent selection, which explains its relative disappearance from the variant landscape over time as vaccination rates have increased. The precise mechanism underlying this behavior is unclear, but several authors have recently shown that S:N501Y exerts epistatic effects on other mutations by altering their antibody escape properties (Starr et al., 2022; Zahradník et al., 2021; Bate et al., 2021). We hypothesize that S:N501Y serves as a linchpin residue within the RBD that constrains the possibilities for vaccine escape when present.

	Alpha PIP	Alpha	Beta PIP	Beta
S:N501Y	0.6056	-0.149 ± 0.266	1.0000	0.408 ± 0.121
ORF1a:A2554V	0.4035	-0.080 ± 0.200	0.5487	-0.066 ± 0.122

Table 4. We report PIPs and estimated coefficients α for vaccination-dependent effects for an analysis in which the vaccination status is 'fully-vaccinated'. We report effects that are ranked in the top 75 by PIP. Estimated β coefficients and PIPs for the corresponding vaccination-independent effects are reported in the two rightmost columns.

4.7. Epistasis

Mounting experimental evidence for epistasis in SARS-CoV-2 (Starr et al., 2022; Javanmardi et al., 2022) raises the question whether these kinds of effects can be inferred from genomic surveillance data. Doing so is expected to be difficult a priori due to the combinatorially large space of possible interaction effects coupled with the fact that many combinations of mutations are unobserved in available data. In this context we expect that explicit sparsity assumptions and high-fidelity statistical inference—key selling points of BVAS—are likely to be crucial.

Here we report initial results of such an analysis, focusing on pairwise interaction effects between non-synonymous mutations in the Spike protein. In particular we consider pairwise interactions between the 421 S mutations in our main analysis—after excluding D614G because it became fixed early in the pandemic—which corresponds to 88410 pairwise interactions. We further exclude pairs of mutations that are not observed together in at least two SARS-CoV-2 clusters—yielding 1432 pairwise interactions—and use BVAS to jointly infer selection effects for the resulting 2975 linear and 1432 pairwise effects. See Sec. S14.8 for details.

We summarize our results as a pair of 'interaction networks' in Figure 11 and report more detailed results in Table S4. We find that top-scoring interaction effects in our analysis are enriched for interactions between mutations that correspond to top-scoring (linear) effects in our default analysis ($P \leqslant 10^{-4}$). Put differently, we find limited evidence for interaction effects between pairs of mutations where one or both mutations are inferred as approximately neutral on their own.

The interaction effect with the largest PIP is between N501Y in the RBD and P681H adjacent to the furin cleavage site. These two mutations appear together in BA.1/B.1.1.7, while N501Y appears without P681H in P.1/B.1.351 and P681H appears without N501Y in B.1.1.159/B.1.243 (Table S4). As such there is a plausible evidential basis for inferring the interaction between N501Y and P681H, with each combination of these amino acids appearing in at least 60k SARS-CoV-2 sequences in our dataset. The importance of N501Y in epistatic interactions is concordant with data from Starr et al. (2022), who found that N501Y caused the largest shifts in the effects of mutations at other sites using deep mutational scanning libraries of RBD and measuring the impact of every amino acid mutation on ACE2-binding affinity.

Among subleading hits, we find a less well resolved picture due to linkage between different pairs of amino acid mutations,

which is reflected in the more moderate PIP scores. For example, all SARS-CoV-2 clusters in our data that carry the pair of amino acids (N501Y, A67V) also carry the pairs (N501Y, P681H) and (A67V, E484A)—but not vice versa. This motivates the graphical representation in Figure 11, which reflects the multiplicity of several amino acid mutations in top-scoring interaction effects—in particular N501Y, P681H, A67V, and L452R.

We expect that resolving these putative interactions in greater detail necessitates follow-up in the lab. Mechanisms of epistasis are likely pleiotropic, but some interactions likely arise from a need to a) incorporate mutations that confer immune escape against the current landscape of circulating variants while b) maintaining protein structure and function. This behavior has been demonstrated for the recurrent S:69-70 deletion (Gupta et al., 2021). Because the space of possible combinations of alleles is very large, most studies have characterized mutations either individually or in the combinations present in common lineages. By ranking combinations of mutations according to evidence of their selective effects, our model may help focus experimental design in this challenging combinatorial setting.

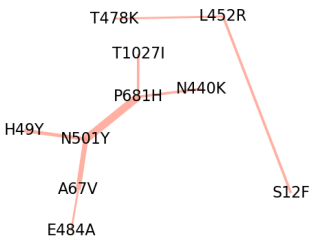

Fig. 11. We depict the two interaction networks of pairwise selection effects in Spike inferred by BVAS. Edge widths are proportional to the posterior inclusion probability of the corresponding pairwise effect. Spatial orientation has no meaning.

5. Discussion

Bayesian Viral Allele Selection unifies and improves upon the two available methods for inferring selection from large scale genomic surveillance data. One of its strengths is that it makes clear assumptions: i) most alleles are neutral; and ii) viral dynamics is governed by an intuitive discrete time branching process. Other advantages of BVAS include its robustness to hyperparameter choices, its satisfactory uncertainty estimates and the fact that it offers Posterior Inclusion Probabilities. Moreover, the diffusion-based likelihood in Eqn. 9 is robust to a number of sources of possible bias, including varying sampling rates across time and space and changes in fitness that affect all lineages equally (due to e.g. lockdown measures or variable temperature/humidity).

We highlight the following limitations.[11] Estimating the effective population size ν is challenging, especially since ν can exhibit significant variability across time. While we have argued that sensitivity to ν is fairly moderate, improved ν estimates should lead to improved statistical efficiency, especially if ν can be estimated with finer spatial and temporal granularity. Doing so would likely require incorporating additional sources of data (e.g. case counts) and represents an important direction for future work. Several of the simplifying assumptions that underly BVAS are expected to be violated at some level in real world data. Notably, our basic fitness model is unable to account for epistasis (see Eqn. 2). While we have extended this linear model to include pairwise interactions, we haved limited this analysis to the Spike protein. Because a genome wide epistasis analysis must contend with millions of possible interactions a priori, additional assumptions to reduce the space of selection effects considered are likely required to make this kind of analysis statistically and computationally tractable. For example, one might limit the analysis to pairs of mutations that are near each other in space.

In summary, BVAS provides a principled statistical and computational framework to identify selection under the constraint of sparsity. Applying BVAS to 6.9 million SARS-CoV-2 genomes provides a detailed picture of viral selection in action. We anticipate that BVAS will be widely applicable to SARS-CoV-2 and other viruses as large scale genomic surveillance data become increasingly available.

6. Data Availability Statement

The SARS-CoV-2 data used in our analysis are provided by GISAID (Elbe and Buckland-Merrett, 2017).[12] A complete list of accession numbers for the viral genomes used in our study is publicly available:
`https://github.com/broadinstitute/bvas/raw/main/paper/accession_ids.txt.xz`
The UShER tree used in our pre-processing pipeline is publicly available: `https://hgwdev.gi.ucsc.edu/~angie/9f94a7b/`. The vaccination data we use are provided by OWID (Ritchie et al., 2020): `https://github.com/owid/covid-19-data/`. An open-source implementation of our analysis code is available at `https://github.com/broadinstitute/bvas`. The initial portion of our data pre-processing pipeline relies on open source code described by Obermeyer et al. (2022): `https://github.com/broadinstitute/pyro-cov`. Allele-level and lineage-level inference results from our main BVAS analysis are publicly available:
`https://github.com/broadinstitute/bvas/raw/main/paper/allele_summary.csv`
`https://github.com/broadinstitute/bvas/raw/main/paper/growth_rates_summary.csv`

7. Acknowledgments

We gratefully acknowledge colleagues from the originating laboratories responsible for obtaining SARS-CoV-2 specimens. Likewise we gratefully acknowledge colleagues from the submitting laboratories where genetic sequence data were generated and shared via the GISAID initiative. This research would not be possible without their collective efforts; see Sec. 6 for more information on the data used. We warmly thank Angie Hinrichs for providing the UShER tree that forms a key component of our data pre-processing pipeline. This work would not be possible without her gracious assistance. We also thank Nikolaos Barkas, Stephen F. Schaffner, Jesse D. Pyle, Lonya Yurkovetskiy, Matteo Bosso, Daniel J. Park, Mehrtash Babadi, Bronwyn L. MacInnis, Jeremy Luban, and Pardis C. Sabeti for discussions about SARS-CoV-2.

[11] Please refer to Sec. S12 for additional discussion.

[12] `https://www.gisaid.org/`

8. Funding

This work was supported in part by grants from MassCPR Viral Variants Program and CDC BAA 75D30120C09605 (to J.E.L).

9. Conflict of Interest

No competing interest is declared.

References

B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu. Superspreading events in the transmission dynamics of sars-cov-2: Opportunities for interventions and control. *PLoS biology*, 18(11):e3000897, 2020.

N. Bate, C. G. Savva, P. C. Moody, E. A. Brown, J. K. Ball, J. W. Schwabe, J. Sale, and N. Brindle. In vitro evolution predicts emerging cov-2 mutations with high affinity for ace2 and cross-species binding. *BioRxiv*, 2021.

Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, et al. Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *The Lancet infectious diseases*, 20(8):911–919, 2020.

Y. R. Cao, A. Yisimayi, F. Jian, W. Song, T. Xiao, L. Wang, S. Du, J. Wang, Q. Li, X. Chen, P. Wang, Z. Zhang, P. Liu, R. An, X. Hao, Y. Wang, J. Wang, R. Feng, H. Sun, L. Zhao, W. Zhang, D. Zhao, J. Zheng, L. Yu, C. Li, N. Zhang, R. Wang, X. Niu, S. Yang, X. Song, L. Zheng, Z. Li, Q. Gu, F. Shao, W. Huang, R. Jin, Z. Shen, Y. Wang, X. Wang, J. Xiao, and X. S. Xie. Ba.2.12.1, ba.4 and ba.5 escape antibodies elicited by omicron infection. *bioRxiv*, 2022. doi: 10.1101/2022.04.30.489997. URL `https://www.biorxiv.org/content/early/2022/05/02/2022.04.30.489997`.

H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, I. H. Solomon, H.-H. Kuo, J. Boucau, K. Bowman, et al. Persistence and evolution of sars-cov-2 in an immunocompromised host. *New England Journal of Medicine*, 383(23):2291–2293, 2020.

P. Colson, P.-E. Fournier, J. Delerce, M. Million, M. Bedotto, L. Houhamdi, N. Yahi, J. Bayette, A. Levasseur, J. Fantini, et al. Culture and identification of a "deltamicron" sars-cov-2 in a three cases cluster in southern france. *Journal of Medical Virology*, 2022.

N. G. Davies, C. I. Jarvis, W. J. Edmunds, N. P. Jewell, K. Diaz-Ordaz, and R. H. Keogh. Increased mortality in community-tested cases of sars-cov-2 lineage b. 1.1. 7. *Nature*, 593(7858): 270–274, 2021.

X. Deng, M. A. Garcia-Knight, M. M. Khalid, V. Servellita, C. Wang, M. K. Morris, A. Sotomayor-González, D. R. Glasner, K. R. Reyes, A. S. Gliwa, et al. Transmission, infectivity, and neutralization of a spike l452r sars-cov-2 variant. *Cell*, 184(13):3426–3437, 2021.

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

A. Endo et al. Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome open research*, 5, 2020.

A. Ferrer-Admetlla, C. Leuenberger, J. D. Jensen, and D. Wegmann. An approximate markov model for the wright–fisher diffusion and its application to time series data. *Genetics*, 203(2):831–846, 2016.

A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. Crawford, et al. Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition. *Cell host & microbe*, 29(1):44–57, 2021.

A. J. Greaney, T. N. Starr, and J. D. Bloom. An antibody-escape estimator for mutations to the sars-cov-2 receptor-binding domain. *Virus Evolution*, 2022.

R. Gupta, S. Kemp, W. Harvey, S. Lytras, A. Carabelli, and D. Robertson. Recurrent independent emergence and transmission of sars-cov-2 spike amino acid h69/v70 deletions. 2021.

Y. Huang, C. Yang, X.-f. Xu, W. Xu, and S.-w. Liu. Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19. *Acta Pharmacologica Sinica*, 41(9):1141–1149, 2020.

S. Iketani, L. Liu, Y. Guo, L. Liu, J. F.-W. Chan, Y. Huang, M. Wang, Y. Luo, J. Yu, H. Chu, et al. Antibody evasion properties of sars-cov-2 omicron sublineages. *Nature*, pages 1–4, 2022.

B. Jackson, M. F. Boni, M. J. Bull, A. Colleran, R. M. Colquhoun, A. C. Darby, S. Haldenby, V. Hill, A. Lucaci, J. T. McCrone, et al. Generation and transmission of interlineage recombinants in the sars-cov-2 pandemic. *Cell*, 184(20): 5179–5188, 2021.

K. Javanmardi, T. H. Segall-Shapiro, C.-W. Chou, D. R. Boutz, R. J. Olsen, X. Xie, H. Xia, P.-Y. Shi, C. D. Johnson, A. Annapareddy, et al. Antibody escape and cryptic cross-domain stabilization in the sars-cov-2 omicron spike protein. *bioRxiv*, 2022.

K. Khan, F. Karim, Y. Ganga, M. Bernstein, Z. Jule, K. Reedoy, S. Cele, G. Lustig, D. Amoako, N. Wolter, N. Samsunder, A. Sivro, J. E. San, J. Giandhari, H. Tegally, S. Pillay, Y. Naidoo, M. Mazibuko, Y. Miya, N. Ngcobo, N. Manickchund, N. Magula, Q. A. Karim, A. von Gottberg, S. S. Abdool Karim, W. Hanekom, B. I. Gosnell, C.-K. Team, R. J. Lessells, T. de Oliveira, M.-Y. S. Moosa, and A. Sigal. Omicron sub-lineages ba.4/ba.5 escape ba.1 infection elicited neutralizing immunity. *medRxiv*, 2022. doi: 10.1101/2022.04.29.22274477. URL `https://www.medrxiv.org/content/early/2022/05/01/2022.04.29.22274477`.

M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

K. E. Kistler and T. Bedford. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses oc43 and 229e. *eLife*, 10:e64509, jan 2021. ISSN 2050-084X. doi: 10.7554/eLife.64509. URL `https://doi.org/10.7554/eLife.64509`.

M. Lacerda and C. Seoighe. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, 198(3):1237–1250, 2014.

M. S. Lau, B. Grenfell, M. Thomas, M. Bryan, K. Nelson, and B. Lopman. Characterizing superspreading events and age-specific infectiousness of sars-cov-2 transmission in georgia, usa. *Proceedings of the National Academy of Sciences*, 117 (36):22430–22435, 2020.

B. Lee, M. S. Sohail, E. Finney, S. F. Ahmed, A. A. Quadeer, M. R. McKay, and J. P. Barton. Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *medRxiv*, pages 2021–12, 2022.

Q. Li, J. Wu, J. Nie, L. Zhang, H. Hao, S. Liu, C. Zhao, Q. Zhang, H. Liu, L. Nie, et al. The impact of mutations in sars-cov-2 spike on viral infectivity and antigenicity. *Cell*, 182 (5):1284–1294, 2020.

M. J. Lista, H. Winstone, H. D. Wilson, A. Dyer, S. Pickering, R. P. Galao, G. De Lorenzo, V. M. Cowton, W. Furnon, N. Suarez, et al. The p681h mutation in the spike glycoprotein confers type i interferon resistance in the sars-cov-2 alpha (b. 1.1. 7) variant. *bioRxiv*, 2021.

Z. Liu, L. A. VanBlargan, L.-M. Bloyet, P. W. Rothlauf, R. E. Chen, S. Stumpf, H. Zhao, J. M. Errico, E. S. Theel, M. J. Liebeskind, et al. Identification of sars-cov-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell host & microbe*, 29(3):477–488, 2021.

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

O. A. MacLean, R. J. Orton, J. B. Singer, and D. L. Robertson. No evidence for distinct types in the evolution of sars-cov-2. *Virus Evolution*, 6(1):veaa034, 2020.

J. McBroome, B. Thornlow, A. S. Hinrichs, A. Kramer, N. De Maio, N. Goldman, D. Haussler, R. Corbett-Detig, and Y. Turakhia. A daily-updated database and tools for comprehensive sars-cov-2 mutation-annotated trees. *Molecular biology and evolution*, 38(12):5819–5824, 2021.

D. Miller, M. A. Martin, N. Harel, O. Tirosh, T. Kustin, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov, et al. Full genome viral sequences inform patterns of sars-cov-2 spread into and within israel. *Nature communications*, 11(1):1–10, 2020.

A. Mohammad, J. Abubaker, and F. Al-Mulla. Structural modelling of sars-cov-2 alpha variant (b. 1.1. 7) suggests enhanced furin binding and infectivity. *Virus Research*, 303: 198522, 2021.

B. Morel, P. Barbera, L. Czech, B. Bettisworth, L. Hübner, S. Lutteropp, D. Serdari, E.-G. Kostaki, I. Mamais, A. M. Kozlov, et al. Phylogenetic analysis of sars-cov-2 data is difficult. *Molecular biology and evolution*, 38(5):1777–1791, 2021.

F. Obermeyer, M. Jankowiak, N. Barkas, S. F. Schaffner, J. D. Pyle, L. Yurkovetskiy, M. Bosso, D. J. Park, M. Babadi, B. L. MacInnis, J. Luban, P. C. Sabeti, and J. E. Lemieux. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science*, 2022. doi: 10.1126/science.abm1208. URL `https://www.science.org/doi/abs/10.1126/science.abm1208`.

O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540–550, 2009.

A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11):1403–1407, 2020.

H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

A. Saito, T. Irie, R. Suzuki, T. Maemura, H. Nasser, K. Uriu, Y. Kosugi, K. Shirakawa, K. Sadamasu, I. Kimura, et al. Enhanced fusogenicity and pathogenicity of sars-cov-2 delta p681r mutation. *Nature*, 602(7896):300–306, 2022.

M. S. Sohail, R. H. Louie, M. R. McKay, and J. P. Barton. Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature biotechnology*, 39(4):472–479, 2021.

T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310, 2020.

T. N. Starr, A. J. Greaney, W. W. Hannon, A. N. Loes, K. Hauser, J. R. Dillen, E. Ferri, A. G. Farrell, B. Dadonaite, M. McCallum, et al. Shifting mutational constraints in the sars-cov-2 receptor-binding domain during viral evolution. *BioRxiv*, 2022.

A. M. Syed, T. Y. Taha, T. Tabata, I. P. Chen, A. Ciling, M. M. Khalid, B. Sreekumar, P.-Y. Chen, J. M. Hayashi, K. M. Soczek, et al. Rapid assessment of sars-cov-2–evolved variants using virus-like particles. *Science*, 374(6575):1626–1632, 2021.

H. Tegally, M. Moir, J. Everatt, M. Giovanetti, C. Scheepers, E. Wilkinson, K. Subramoney, S. Moyo, D. G. Amoako, C. L. Althaus, U. J. Anyaneji, D. Kekana, R. Viana, J. Giandhari, T. G. Maponga, D. Maruapula, W. Choga, S. H. Mayaphi, N. Mbhele, S. Gaseitsiwe, N. Msomi, Y. Naidoo, S. Pillay, T. a. Sanko, J. E. San, L. Scott, L. Singh, N. A. Magini, P. Smith-Lawrence, W. S. Stevens, G. Dor, D. Tshiabuila, N. Wolter, W. Preiser, F. K. Treurnicht, M. Venter, M. Davids, G. Chiloane, A. Mendes, C. McIntyre, A. O'Toole, C. Ruis, T. P. Peacock, C. Roemer, C. Williamson, O. G. Pybus, J. N. Bhiman, A. J. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, C. Baxter, R. J. Lessells, and T. de Oliveira. Continued emergence and evolution of omicron in south africa: New ba.4 and ba.5 lineages. *medRxiv*, 2022. doi: 10.1101/2022.05.01.22274406.

J. Terhorst, C. Schlötterer, and Y. S. Song. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS genetics*, 11(4):e1005069, 2015.

Y. Turakhia, B. Thornlow, A. S. Hinrichs, N. De Maio, L. Gozashti, R. Lanfear, D. Haussler, and R. Corbett-Detig. Ultrafast sample placement on existing trees (usher) enables real-time phylogenetics for the sars-cov-2 pandemic. *Nature Genetics*, 53(6):809–816, 2021.

D. VanInsberghe, A. S. Neish, A. C. Lowen, and K. Koelle. Recombinant sars-cov-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evolution*, 7(2):veab059, 2021.

E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, et al. Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england. *Nature*, 593(7858):266–269, 2021.

Y. Weisblum, F. Schmidt, F. Zhang, J. DaSilva, D. Poston, J. C. Lorenzi, F. Muecksch, M. Rutkowska, H.-H. Hoffmann, E. Michailidis, et al. Escape from neutralizing antibodies by sars-cov-2 spike protein variants. *Elife*, 9:e61312, 2020.

X. Xie, Y. Cao, A. Yisimayi, F. Jian, W. Song, T. Xiao, L. Wang, S. Du, j. wang, Q. Li, et al. Ba. 2.12. 1, ba. 4 and ba. 5 escape antibodies elicited by omicron ba. 1 infection. 2022.

L. Yurkovetskiy, X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, et al. Structural and functional analysis of the d614g sars-cov-2 spike protein variant. *Cell*, 183(3):739–751, 2020.

J. Zahradník, S. Marciano, M. Shemesh, E. Zoler, D. Harari, J. Chiaravalli, B. Meyer, Y. Rudich, C. Li, I. Marton, et al. Sars-cov-2 variant prediction and antiviral drug design are enabled by rbd in vitro evolution. *Nature microbiology*, 6(9): 1188–1198, 2021.

G. Zanella and G. Roberts. Scalable importance tempering and bayesian variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):489–517, 2019.

T. Zhou, Y. Tsybovsky, J. Gorman, M. Rapp, G. Cerutti, G.-Y. Chuang, P. S. Katsamba, J. M. Sampson, A. Schön, J. Bimela, et al. Cryo-em structures of sars-cov-2 spike without and with ace2 reveal a ph-dependent switch to mediate endosomal positioning of receptor-binding domains. *Cell host & microbe*, 28(6):867–879, 2020.