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

Martin Jankowiak, Fritz H. Obermeyer, Jacob E. Lemieux

This supplement is organized as follows. In Sec. S1 we describe the discrete-time branching process and diffusion approximation that forms the basis of our modeling approach. In Sec. S2 we summarize the simplest approach to statistical inference that utilizes the resulting diffusion-based likelihood, namely maximum a posteriori (MAP) inference. In Sec. S3 we describe our approach to inducing sparsity in the same class of diffusion-based models via Bayesian Variable Selection. In Sec. S4 we give a high-level overview of how we do Markov Chain Monte Carlo (MCMC) inference in this class of models. In Sec. S5 we generalize the discussion to the case of multiple spatial regions. In Sec. S6 we introduce the effective population size estimators we use. In Sec. S7 we generalize our basic model to include vaccination-dependent selection effects. In Sec. S8 we discuss the effect of sampling, i.e. the fact that not all viral sequences are observed. In Sec. S9 we provide an extended discussion of the prior inclusion probability \( h \), which is one of the main hyperparameters of our method. In Sec. S10 we discuss alternative approaches to modeling and inference and why we prefer Bayesian Viral Allele Selection. In Sec. S11 we briefly discuss how we report uncertainty estimates. In Sec. S12 we discuss some of the limitations of BVAS. In Sec. S13 we describe how we simulate viral dynamics and present additional experiments not included in the main text. In Sec. S14 we give additional details and report additional results for our SARS-CoV-2 analysis, including inference results using data obtained through August 10\textsuperscript{th} 2022. In Sec. S15 we provide some references to literature about Bayesian variable selection.

S1 From discrete-time branching process to diffusion

We provide an overview of the population genetics based approach we take to formulating a likelihood that connects observed count data to unobserved selection coefficients. For additional details (in somewhat different notation) please refer to [1].

S1.1 Negative Binomial branching process

A natural model of viral infection is to suppose that an infected individual at time step \( t \) infects a stochastic number of individuals in the next time step, where the number of infected individuals is governed by a Negative Binomial distribution with mean \( \bar{R} \) and variance \( \bar{R} + \bar{R}^2/k \), where \( k \) is the dispersion parameter. In the limit that \( k \to \infty \) we recover a Poisson distribution with mean and variance both equal to \( \bar{R} \). In the opposite ‘super-spreading’ limit, \( k \to 0 \), the Negative Binomial distribution has a (potentially very large) variance that scales like \( k^{-1} \). We now suppose that there are \( n \) infected individuals at time step \( t \). Under this assumption the total number of infected individuals at time step \( t+1 \) is governed by a Negative Binomial distribution with mean \( n\bar{R} \) and variance given by 

\[
(n\bar{R}) + (n\bar{R})^2/(nk) = n\bar{R} + n\bar{R}^2/k.
\]

To make the model more realistic we suppose there are \( V \) distinct viral variants in circulation. Each viral variant \( v \) has an effective reproduction number given by \( R_v = R_0(1 + \Delta R_v) \) where \( R_0 \) corresponds to the wild type. At the current time step \( t \) there are \( n_v(t) \) individuals infected with variant \( v \). Given our assumptions, the number of infected individuals with variant \( v \) at the next time step is governed by the distribution

\[
n_v(t+1) \sim \text{NegativeBinomial}(\text{mean} = n_v(t)R_v, \text{dispersion} = n_v(t)k)
\]

where we have assumed that \( k \) is the same for all \( V \) variants and disregard the possibility of mutation.
S1.2 Diffusion in variant frequency space

Following \cite{1} we now transform from case counts \( n_v(t) \) to variant frequencies \( z_v(t) \):

\[
z_v(t) = \frac{n_v(t)}{\sum_{v'} n_{v'}(t)} \quad (2)
\]

In the limit of large population size, i.e. the diffusion limit \( n \gg 1 \), and under the assumption that \( |\Delta R_v| \ll 1 \), \cite{1} compute approximate first and second moments for the variant frequencies \( z_v(t) \) under the dynamics in Eqn. \cite{1}. These moments can in turn be used to define an (approximately) equivalent diffusion process for continuous-valued variant frequencies \( z_v(t) \in [0,1] \). In particular one finds that \( z(t) \) is approximately distributed like a Multivariate-Normal random variable as

\[
z(t + 1) \sim \mathcal{N}(z(t) + \tilde{d}(t), \nu^{-1} \tilde{\Lambda}(t)) \quad (3)
\]

where \( \tilde{d}(t) \in \mathbb{R}^V \) is the \( V \)-dimensional drift

\[
\tilde{d}_v(t) = z_v(t) \left( \Delta R_v - \sum_{w} \Delta R_{w} z_w(t) \right) \quad (4)
\]

and \( \tilde{\Lambda}(t) \) is the \( V \times V \) (unscaled) diffusion matrix given by

\[
\tilde{\Lambda}_{v,w}(t) = \begin{cases} 
  z_v(t)(1 - z_v(t)) & \text{if } v = w \\
  -z_v(t)z_w(t) & \text{if } v \neq w 
\end{cases} \quad (5)
\]

and \( \nu \) is the effective population size given by

\[
\nu \equiv \left( \frac{1}{R_0} + \frac{1}{k} \right)^{-1} n = \frac{k R_0}{k + R_0} n \quad (6)
\]

Eqn. \cite{3} has a simple and intuitive form with the properties we would expect. For example, the drift \( \tilde{d}(t) \) equals zero and thus the mean of \( z(t + 1) \) is equal to \( z(t) \) if all variants have the same reproduction number (i.e. \( \Delta R_v = 0 \) for all \( v \)). Similarly, the variant \( v \) with the largest \( \Delta R_v \) will tend to see its prevalence increase, i.e. \( \mathbb{E}[z_v(t + 1) - z_v(t)] > 0 \). For fixed \( n \) the effective population size \( \nu \) decreases as \( k \) decreases, i.e. as the dynamics become increasingly characterized by super-spreading and are thus increasingly stochastic. In particular large populations experience approximately deterministic dynamics dominated by drift, while small populations experience considerably more stochastic dynamics in which the drift is more difficult to discern.

S1.3 Diffusion in allele frequency space

Following \cite{1}, it is convenient to transform to a representation based on allele frequencies so that we can model allele-level selection. Suppose the virus has \( A \) alleles with time-dependent allele frequencies \( 0 \leq x_a(t) \leq 1 \) for \( a = 1, \ldots, A \), i.e. \( x_a(t) \) is the fraction of infected individuals at time \( t \) infected by a virus that carries allele \( a \). Denote pairwise allele frequencies as \( x_{ab}(t) \), i.e. \( x_{ab}(t) \) is the fraction of infected individuals at time \( t \) infected by a virus that carries both allele \( a \) and allele \( b \). For convenience we define \( x_{aa}(t) = x_a(t) \). Further suppose that each variant \( v \) is given as a genotype vector \( g_v \in \{0,1\}^A \), i.e. \( g_{v,a} \) encodes whether variant \( v \) carries allele \( a \) (\( g_{v,a} = 1 \)) or not (\( g_{v,a} = 0 \)). We also assume that the variant-level differential reproduction number \( \Delta R_v \) is governed by a linear additive model in allele-space:

\[
\Delta R_v = \sum_a g_{v,a} \beta_a \quad (7)
\]

where \( \beta_a \in \mathbb{R} \) is an allele-level selection coefficient with positive (respectively, negative) coefficients corresponding to increased (decreased) viral fitness.

With these ingredients in hand and under the linear assumption in Eqn. \cite{7} we can now specify the dynamics in allele frequency space, where the dynamics in Eqn. \cite{3} are transformed to

\[
x(t + 1) \sim \mathcal{N}(x(t) + d(t), \nu^{-1} \Lambda(t)) \quad (8)
\]

\[
\Delta R_v = \sum_a g_{v,a} \beta_a \quad (7)
\]

where \( \beta_a \in \mathbb{R} \) is an allele-level selection coefficient with positive (respectively, negative) coefficients corresponding to increased (decreased) viral fitness.

With these ingredients in hand and under the linear assumption in Eqn. \cite{7} we can now specify the dynamics in allele frequency space, where the dynamics in Eqn. \cite{3} are transformed to

\[
x(t + 1) \sim \mathcal{N}(x(t) + d(t), \nu^{-1} \Lambda(t)) \quad (8)
\]
where \( \mathbf{d}(t) \in \mathbb{R}^A \) is the \( A \)-dimensional drift given by
\[
\mathbf{d}(t) = \mathbf{A}(t)\mathbf{\beta} \quad \iff \quad d_a(t) = x_a(t)(1 - x_a(t))\beta_a + \sum_{b \neq a} (x_{ab}(t) - x_a(t)x_b(t))\beta_b \tag{9}
\]
and \( \nu \) is as in Eqn. 6. Here \( \mathbf{A}(t) \), analog to the diffusion matrix in Eqn. 5, is given by the \( A \times A \) matrix
\[
\mathbf{A}_{ab}(t) = x_{ab}(t) - x_a(t)x_b(t) \tag{10}
\]
Note that whenever alleles \( a \) and \( b \) exhibit non-trivial correlation, i.e. whenever \( x_{ab}(t) \neq x_a(t)x_b(t) \), \( d_a(t) \) depends on \( \beta_b \) (in addition to \( \beta_a \)). In other words, a neutral allele can be ‘dragged along’ (i.e. exhibit apparent drift) by a non-neutral allele via hitchhiking. This effect is sometimes called genetic draft.

It is convenient to introduce incremental allele frequency changes
\[
\mathbf{y}(t) = \mathbf{x}(t+1) - \mathbf{x}(t) \tag{11}
\]
so that Eqn. 8 becomes
\[
\mathbf{y}(t) \sim \mathcal{N}(\mathbf{d}(t), \nu^{-1}\mathbf{A}(t)) \tag{12}
\]
Eqn. 12 together with the definitions in Eqn. 9-11 forms the basis of everything that follows.

### S1.4 Relationship to PyR_0

An alternative approach to inferring allele-level selection effects is the PyR_0 model described in [2]. At its core this hierarchical Bayesian model relies on the following multivariate logistic growth ansatz
\[
z_v(t) = \frac{e^{\Delta \mathbf{R}_v t}}{\sum_v e^{\Delta \mathbf{R}_v t}} \tag{13}
\]
where \( \mathbf{z}(t) \) are variant frequencies like in Eqn. 3 and \( \Delta \mathbf{R}_v \) are variant-level differential growth rates. In direct analogy to the approach adopted in this work as well as [1] PyR_0 proceeds to regress \( \Delta \mathbf{R}_v \) against allele-level features (cf. Eqn. 7). How is the PyR_0 ansatz in Eqn. 13 related to the diffusion-based dynamics in Eqn. 3? Eqn. 3 can be viewed as a specification of a stochastic differential equation in which there is deterministic drift \( \mathbf{d}(t) \) as well as stochastic diffusion controlled by \( \mathbf{A}(t) \). In the infinite population limit, \( \nu \to \infty \), the dynamics are deterministic and reduce to an ordinary differential equation. It is easy to check that Eqn. 13 is a solution to precisely that differential equation, namely
\[
\frac{d}{dt}\mathbf{z}(t) = \tilde{\mathbf{d}}(t) \tag{14}
\]
It would thus appear that the PyR_0 likelihood can be viewed as the deterministic counterpart of the diffusion-based likelihood in Eqn. 3. However, this is not quite correct, since PyR_0 uses the ansatz in Eqn. 13 to parameterize the probabilities of a \( V \)-dimensional Multinomial distribution. That is, PyR_0 describes an infinite population undergoing deterministic dynamics, with only a finite number of viral sequences observed at each time point. As a consequence the noise structure of the PyR_0 Multinomial likelihood mirrors that of Eqn. 3 with the difference that the Multinomial likelihood implicitly assumes a fixed population size of \( n \), where \( n \) is the total number of observed viral sequences at a given time step. Importantly, it does not allow for the reduction in effective population size that results from super-spreading (small \( k \), see Eqn. 6) or a finite sampling rate (see Sec. S5).

As we will see in our analysis of SARS-CoV-2 data, the effective population size in real data is quite modest, even in relatively well-sampled regions. Thus the stochastic component of the dynamics is expected to play an important role and, we would argue, should be accounted for explicitly in the model using a data-driven estimate of the effective population size. The ability to do so is one of the main benefits of using a diffusion-based likelihood.

---

1. Note that Sec. S1.2 and Sec. S1.4 are the only sections in the supplement where we consider diffusion in variant-frequency space. In particular elsewhere drift vectors and diffusion matrices always correspond to diffusion in allele-frequency space.
2. Note that a Multinomial distribution with \( n \) trials and event probabilities \( p_v \) for \( v = 1, ..., V \) has variance \( np_v(1 - p_v) \) and covariance \(-np_v p_w \) (for \( v \neq w \)).
S2  Maximum a posteriori inference (MAP)

The simplest model that utilizes the diffusion-based likelihood in Eqn. 12 is formulated as follows (we refer the reader to [1] for additional discussion). First we place a Multivariate-Normal prior on the $A$-dimensional vector of selection coefficients $\beta$

$$p(\beta|\nu) = \mathcal{N}(\beta|0, \nu^{-1}\mathbb{1}_A)$$

(15)

where $\nu > 0$ is the prior precision on the coefficients and $\mathbb{1}_A$ is the $A \times A$ identity matrix. For observed incremental allele frequency changes $y(t)$ for $t = 1, ..., T-1$ the likelihood is given by

$$p(y_{1:T-1}|\beta, \nu) = \prod_{t=1}^{T-1} \mathcal{N}(d(t|\beta), \nu^{-1}\Lambda(t))$$

(16)

where we have assumed that $\nu$ is constant across time. Since $\beta$ appears linearly in the drift $d(t|\beta)$ (see Eqn. 9) and the prior is Multivariate-Normal, the corresponding posterior distribution, which is given by

$$p(\beta|y_{1:T-1}, \nu, \tau) = \frac{\mathcal{N}(\beta|0, \tau^{-1}\mathbb{1}_A) \prod_{t=1}^{T-1} \mathcal{N}(d(t|\beta), \nu^{-1}\Lambda(t))}{\int d\beta \mathcal{N}(\beta|0, \tau^{-1}\mathbb{1}_A) \prod_{t=1}^{T-1} \mathcal{N}(d(t|\beta), \nu^{-1}\Lambda(t))}$$

(17)

can be computed in closed form and is itself a Multivariate-Normal distribution. In particular the posterior mean, which in this case is also the maximum a posteriori (MAP) estimate, is given by

$$\beta^{\text{MAP}} = \arg \max_{\beta} p(\beta|y_{1:T-1}|\beta, \nu)$$

(18)

$$= \mathbb{E}_{p(\beta|y_{1:T-1}, \nu, \tau)}[\beta]$$

(19)

$$= \left( \sum_{t=1}^{T-1} \Lambda(t) + \frac{\tau}{\nu} \mathbb{1}_A \right)^{-1} \sum_{t=1}^{T-1} y(t)$$

(20)

$$= \left( \sum_{t=1}^{T-1} \Lambda(t) + \frac{\tau}{\nu} \mathbb{1}_A \right)^{-1} \left( x(T) - x(1) \right)$$

(21)

We note that the effective regularization parameter for MAP is given by the ratio between the prior precision $\tau$ and the effective population size $\nu$:  

$$\gamma_{\text{reg}} = \frac{\tau}{\nu}$$

(22)

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 hardware, at least for $A$ up to $A \sim 10^3 - 10^5$. An unattractive property of this estimator is that it can perform poorly in the high-dimensional case, $A \gg 1$, since we expect most alleles to be neutral, but $\beta^{\text{MAP}}$ will generally be non-zero for all $a$.

S3  Bayesian Viral Allele Selection

We expect most alleles to be neutral ($\beta_a = 0$) or 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 [3]. In more detail we consider the following space of models:

- inclusion variables $\gamma_a \sim \text{Bernoulli}(h)$ for $a = 1, ..., A$
- selection coefficients $\beta \gamma \sim \mathcal{N}(0, \tau^{-1}\mathbb{1}_{|\gamma|})$
- allele frequency changes $y(t) \sim \mathcal{N}(d(t|\beta_s), \nu^{-1}\Lambda(t))$ for $t = 1, ..., T-1$

Here each Bernoulli latent variable $\gamma_a \in \{0, 1\}$ controls whether the $a^{th}$ coefficient $\beta_a$ is included ($\gamma_a = 1$) or excluded ($\gamma_a = 0$) from the model; in other words it controls whether the $a^{th}$ allele is neutral ($\gamma_a = 0$) or not ($\gamma_a = 1$). In the following we use $\gamma$ to refer to the full $A$-dimensional vector ($\gamma_1, ..., \gamma_A$). The hyperparameter $h \in (0, 1)$ controls the overall level of sparsity; in particular $hA$ is the expected number of non-neutral alleles a priori. The $|\gamma|$ coefficients $\beta_s \in \mathbb{R}^{|\gamma|}$ are governed by a

\[^3\text{For more discussion see Sec. S9 below.}\]
Normal prior with precision \( \tau \) where \( \tau > 0 \) is a fixed hyperparameter. Here \( \gamma \in \{0, 1, \ldots, A\} \) denotes the total number of non-neutral alleles in a given model. Finally we note that the drift \( d(t|\beta_b) \) is given by \( \Lambda(t)\beta_b \), i.e. all alleles \( b \) not included in the model (alleles with \( \gamma_b = 0 \)) can be understood as having null coefficients \( \beta_b = 0 \). In the following we drop the \( \gamma \) subscript on \( \beta_b \) to simplify the notation.

In addition to inducing sparsity, an attractive feature of the model in Eqn. \( 23 \) 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 \( 0 \leq \text{PIP} \leq 1 \). The PIP is defined as

\[
\text{PIP}(a) = p(\gamma_a = 1|y_{1:T-1}, \nu)
\]

i.e. PIP(\( a \)) is the posterior probability that allele \( a \) is included in the model. This quantity should be contrasted to \( h \) in Eqn. \( 23 \) which is the \textit{a priori} inclusion probability. Alleles that have large PIPs are good candidates for being causally linked to viral fitness.

S4 MCMC Inference

In this section we describe our approach to inference for the model in Sec. S3. Before we do so it is worth emphasizing that this is a difficult inference problem. There are two basic reasons for this. First, this is a transdimensional inference problem defined on a mixed discrete/continuous latent space. In particular the dimension of \( \beta \) depends on \( |\gamma| \). Second, the size of the model space is astronomically large whenever there are more than a few dozen alleles. Indeed for \( A \) alleles the total number of distinct models, namely \( 2^A \), exceeds the estimated number of atoms in the known universe (\( \sim 10^{80} \)) for \( A \gtrsim 266 \). However, the model also exhibits conjugacy structure that we can exploit. In particular, as we discuss next, \( \beta \) can be integrated out analytically for any given \( \gamma \).

S4.1 Integrating out \( \beta \)

Recall from Eqn. \( 9 \) that the drift is given by \( d(t) = \Lambda(t)\beta \) so that each term in the likelihood (see Eqn. \( 12 \)) can be written as

\[
(y(t) - d(t))^T\Lambda(t)^{-1}(y(t) - d(t)) = \\
d(t)^T\Lambda(t)^{-1}d(t) - 2d(t)^T\Lambda(t)^{-1}y(t) = \beta^T\Lambda(t)\beta - 2\beta^Ty(t)
\]

(25)

(26)

where we have ignored the pre-factor of \( -\frac{\nu}{2} \) and dropped terms that do not depend on \( \beta \). After re-introducing \( -\frac{\nu}{2} \), each term in our likelihood contains the following quadratic form:

\[
-\frac{\nu}{2} \left( \beta^T\Lambda(t)\beta - 2\beta^Ty(t) \right)
\]

(27)

With this formula in hand we can compute the log marginal likelihood in closed form for any particular \( \gamma \); up to irrelevant constants we have:

\[
\log p(y_{1:T-1}|\gamma) = \frac{1}{2}y^\nu^T \left( \Lambda^\nu + \tau \mathbb{1}_{|\gamma|} \right)^{-1}y^\nu - \frac{1}{2} \log \det \left( \Lambda^\nu + \tau \mathbb{1}_{|\gamma|} \right) + \frac{1}{2} \log \det \left( \tau \mathbb{1}_{|\gamma|} \right)
\]

(28)

where

\[
y^\nu = \nu \sum_{t=1}^{T-1} y(t) = \nu (x(T) - x(1)) \\
\Lambda^\nu = \nu \sum_{t=1}^{T-1} \Lambda(t)
\]

(29)

Eqn. \( 28 \) can be computed efficiently using basic linear algebra tricks described in e.g. \( 4 \) and \( 5 \). It is also important to note that \( y^\nu \) and \( \Lambda^\nu \) only need to be computed once before we run MCMC inference. In particular this means that MCMC iteration cost does not depend on \( T \), since the summation over \( t \) in Eqn. \( 29 \) can be done in pre-processing.

Crucially, since \( \beta \) can be integrated out analytically, our inference problem reduces to an inference problem over \( \gamma \)-space. This is convenient because \( \gamma \in \{0, 1\}^A \) has fixed dimensionality and so we no longer need to concern ourselves with transdimensional inference.
S4.2 MCMC

The resulting inference problem is still difficult, since the posterior over $\gamma$ is defined on a space of size $2^A$. Gibbs sampling offers a simple approach to inference for this model. However, Gibbs sampling mixes exicruciatingly slowly unless $A$ is very small. To obtain an efficient algorithm we utilize a recently introduced MCMC inference algorithm, Tempered Gibbs Sampling, described in [4]. Here we give a very high-level description of this algorithm and refer the reader to [4] for more details.

We emphasize two main features of this algorithm. The first feature is that the MCMC algorithm targets an auxiliary posterior distribution that is tempered w.r.t. our original target distribution. Samples from the original target distribution are then obtained by reweighting, i.e. importance sampling. Crucially the algorithm is designed in such a way that tempering dramatically improves mixing while the variance of the importance weights remains moderate. Second, the algorithm makes extensive use of the fact that Eqn. 28 can be computed in closed form. In particular it makes use of the fact that, for any given value of $\gamma$, it is relatively cheap to compute the quantity $p(\gamma_a|\gamma_{-a}, y_{1:T-1})$ for all $a$ in parallel. Here $\gamma_{-a}$ denotes all the entries of $\gamma$ except for $a$. This information is then used by the algorithm to make greedy moves in the immediate neighborhood of $\gamma$. Consequently the algorithm is able to quickly and efficiently explore regions of $\gamma$ space with high posterior probability.

S4.3 Computational complexity

The computational complexity of a single MCMC iteration is given by $\mathcal{O}(|\gamma|^3 + |\gamma|^2 A)$, where $|\gamma|$ is the number of non-neutral alleles at a given iteration and $A$ is the total number of alleles. If we assume that the sparsity assumption is valid so that $|\gamma| \ll A$ the computational complexity simplifies to $\mathcal{O}(|\gamma|^2 A)$ and is dominated by the cost of computing $p(\gamma_a|\gamma_{-a}, y_{1:T-1})$ for $a = 1, ..., A$. In practice we find that this is quite fast on a mid-grade GPU. For example, for $A = 2904$ and using a NVIDIA Tesla T4 GPU it takes about 45 minutes to generate half a million MCMC samples, which is about 200 samples per second.

S5 Multiple spatial regions

Above we have assumed that observed allele frequencies $x(t)$ are from a single spatial region. The generalization to multiple regions is straightforward. First we add a $r = 1, ..., N_R$ subscript to the likelihood:

$$y_r(t) \sim \mathcal{N}(d_r(t), \nu_r^{-1}\Lambda_r(t)) \quad \text{for} \quad r = 1, ..., N_R \quad \text{and} \quad t = 1, ..., T - 1 \quad (30)$$

Depending on the strategy used to estimate the effective population size (see Sec. S6) each $\nu_r$ in Eqn. 29 is either distinct or is instead a single global value. The equation for the marginal log likelihood, Eqn. 28, is unchanged except we now write Eqn. 29 as

$$\tilde{y}^\nu = \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} y_r(t) \quad \tilde{A}^\nu = \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \Lambda_r(t) \quad (31)$$

This equation makes it evident that our effective population size estimates play an important role, since they control the relative contribution of different regions to the region-summed totals in $\tilde{y}^\nu$ and $\tilde{A}^\nu$. In particular using a per-region $\nu$ estimator has the consequence that regions with larger effective population sizes contribute more to the likelihood than regions with small effective population sizes. This is of course as expected, since regions with smaller effective population sizes are more stochastic and thus contribute less certain information about the drift. We discuss our strategy for estimating $\nu$ in the next section.

S6 Inferring the effective population size $\nu$

The likelihood in Eqn. 12 depends on the effective population size $\nu$, a quantity that we do not know a priori and need to estimate from data. By assumption the increment $y(t)$ is distributed as

$$y(t) \sim \mathcal{N}(d(t), \nu^{-1}\Lambda(t))$$

This information is then used by the algorithm to make greedy moves in the immediate neighborhood of $\gamma$. Consequently the algorithm is able to quickly and efficiently explore regions of $\gamma$ space with high posterior probability.
Under this assumption
\[ \mathbb{E} \left[ y(t)^{T} y(t) \right] = d(t)^{T} d(t) + \nu^{-1} \text{Tr} \Lambda(t) \tag{33} \]
If we assume that the drift term is subdominant so that
\[ \mathbb{E} \left[ y(t)^{T} y(t) \right] \approx \nu^{-1} \text{Tr} \Lambda(t) \tag{34} \]
this leads to the estimator
\[ \hat{\nu} = \frac{\text{Tr} \Lambda(t)}{\mathbb{E} \left[ y(t)^{T} y(t) \right]} \tag{35} \]
We note that, since \( d(t)^{T} d(t) \geq 0 \), we might expect \( \hat{\nu} \) to be an underestimate of \( \nu \).

Thus for a particular region we can estimate the effective population size as
\[ \hat{\nu} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{Tr} \Lambda(t) \left( \frac{1}{\mathbb{E} \left[ y(t)^{T} y(t) \right]} \right) \tag{36} \]
where we have averaged Eqn. 35 over \( T-1 \) time steps. We can also partially correct for the bias introduced by the (a priori unknown and non-zero) drift by instead using a centered estimator:
\[ \hat{\nu} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{Tr} \Lambda(t) \left( \frac{1}{T} \right) \left( \frac{1}{T} \right) \left( x(T) - x(1) \right) \tag{37} \]
with
\[ \bar{y} = \sum_{t=1}^{T-1} y(t) = x(T) - x(1) \tag{38} \]
denoting time-integrated allele increments. We find that both estimators in Eqn. 36 and Eqn. 37 work well in simulation, but for simplicity we use Eqn. 36 throughout this work. Indeed it is also possible to construct \( \hat{\nu} \) estimators that are more time-local in nature and can thus accommodate population sizes that change over time, but for simplicity we prefer to use the hyperparameter-free estimator in Eqn. 36 as we expect this choice to be more robust. More complex schemes are certainly possible, but it is likely that they would require additional sources of data beyond (partially observed) genomic surveillance data if they are to be reliable (e.g. case count data).

The discussion above assumes a single spatial region. How can we accommodate multiple regions? We adopt two possible schemes. In the global scheme we compute \( \hat{\nu} \) within each region using Eqn. 36 and then compute a single global effective population size \( \hat{\nu} \) via a mean or median over all regions. In the regional scheme we simply compute \( \hat{\nu} \) within each region using Eqn. 36. Since we expect the latter scheme to lack robustness when confronted with the complexities of real-world SARS-CoV-2 data, we only make comparisons to the regional scheme in simulation.

S7 Incorporating time-dependent vaccination rates

Suppose we know the vaccination rate \( 0 \leq \phi_r(t) \leq 1 \) for a given region \( r \). We would like to incorporate this information into our modeling. In particular we would like to allow for vaccination-dependent selection. To do this we write the drift in region \( r \) as
\[ d_{r,a}(t) = x_{r,a}(t) \left( 1 - x_{r,a}(t) \right) \left( \beta_a + \phi_r(t) \alpha_a \right) + \sum_{b \neq a} \left( x_{r,ab}(t) - x_{r,a}(t)x_{r,b}(t) \right) \left( \beta_b + \phi_r(t) \alpha_b \right) \tag{39} \]
where \( \alpha \in \mathbb{R}^A \) is a second group of selection coefficients whose strength is modulated by the time- and region-local vaccination rate \( \phi_r(t) \). In particular \( \alpha \) only has a non-negligible effect on viral dynamics when the vaccination rate has attained a non-negligible level. Disentangling the effects of \( \beta \) and \( \alpha \) is difficult a priori. However, our hope is that a Bayesian variable selection approach with robust MCMC inference should be up to the task provided we have enough data.

The particular form of Eqn. 39 can be motivated as follows. Eqn. 39 implies that for any given allele there are two selection coefficients: one for unvaccinated populations (\( \beta \)) and another for vaccinated populations (\( \alpha + \beta \)). These population-level effects can be connected back to the individual level. In particular take the \( k \to \infty \) limit of the discrete time process in Sec. S1 so that secondary infections are controlled by a Poisson distribution. Suppose we are considering a
variant with a single non-wild-type allele and there are $n$ infected individuals at the current time step. If we suppose that secondary infections result from contact between infected and non-infected individuals, and that an infected individual comes into contact with non-infected individuals who are vaccinationed/unvaccinated in the proportion $\phi$ to $1-\phi$, then the total number of secondary infections in the next time step is given by

$$n_{\text{unvac}} + n_{\text{vac}} \quad \text{with} \quad n_{\text{unvac}} \sim \text{Poisson}(R_0(1+\beta)(1-\phi)n) \quad \text{and} \quad n_{\text{vac}} \sim \text{Poisson}(R_0(1+\beta+\alpha)\phi n)$$

(40)

which is governed by a Poisson($R_0(1+\beta+\phi\alpha)n$) distribution. Here $n_{\text{unvac}}$ and $n_{\text{vac}}$ are the number of secondary infections among unvaccinated and vaccinated individuals, respectively. Thus the effective selection coefficient is given by $\beta + \phi\alpha$, explaining the form of Eqn. 39.

The basic inference approach introduced in Sec. S4 is still applicable. All we need to do is suitably modify $\mathbf{y}^\nu$ and $\mathbf{A}^\nu$ in Eqn. 31. In particular we now have that $\mathbf{y}^\nu \in \mathbb{R}^{2A}$

$$\mathbf{y}^\nu = \begin{pmatrix} \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \mathbf{y}_r(t) \\ \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \phi_r(t) \mathbf{y}_r(t) \end{pmatrix}$$

(41)

and $\mathbf{A}^\nu \in \mathbb{R}^{2A \times 2A}$:

$$\mathbf{A}^\nu = \begin{pmatrix} \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \mathbf{A}_r(t) & \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \phi_r(t) \mathbf{A}_r(t) \\ \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \phi_r(t) \mathbf{A}_r(t) & \sum_{r=1}^{N_R} \nu_r \sum_{t=1}^{T-1} \phi_r^2(t) \mathbf{A}_r(t) \end{pmatrix}$$

(42)

This change approximately doubles the cost of MCMC inference and quadruples memory requirements, but this regime is still well within reach of a mid-grade GPU even for $A \sim 10^4$.

### S8 Sampling rate

What effect does sampling—the fact that not all viral sequences are observed in real world datasets—have on our diffusion-based likelihood? We first give a somewhat high-level analysis that ignores some of the subtleties involved. Suppose that each infected individual has their viral genome sequenced with probability $0 < \rho < 1$. Due to sampling, a variant $v$ that is circulating in $n_v$ individuals will be observed $\bar{n}_v \sim \text{Binomial}(n_v, \rho)$ times. This distribution has mean $\rho n_v$ and variance $\rho(1-\rho)n_v$. For $\rho \ll 1$ and large $n_v$ we have that $\bar{n}_v$ is very nearly governed by the Poisson distribution $\text{Poisson}(\rho n_v)$, which has mean and variance both equal to $\rho n_v$. Note, however, that the Poisson distribution is just the $k \to \infty$ limit of the Negative Binomial distribution, which in turn is the basic ingredient of the discrete time process that underlies our diffusion-based likelihood (see Eqn. 1). For this reason the combined action of one step of viral infection and sampling rate on the allele frequencies $x_v(t)$ can be described by the equation

$$x(t+1) \sim x(t) + \mathbf{d}(t) + \mathbf{\epsilon} + \mathbf{\epsilon}'$$

(43)

with

$$\mathbf{\epsilon} \sim \mathcal{N}(0, \nu^{-1}\mathbf{A}(t)) \quad \quad \mathbf{\epsilon}' \sim \mathcal{N}(0, \nu'^{-1}\mathbf{A}(t))$$

(44)

where

$$\nu' = \rho n$$

(45)

is the effective population size corresponding to variability due to sampling and where $\mathbf{\epsilon}$ and $\mathbf{\epsilon}'$ encode variability in infection and sampling, respectively. Since, however, both sources of variability are zero mean normally distributed noise in the diffusion limit, they can easily be combined:

$$x(t+1) \sim x(t) + \mathbf{d}(t) + \mathbf{\epsilon}'' \quad \text{with} \quad \mathbf{\epsilon}'' \sim \mathcal{N}(0, \nu''^{-1}\mathbf{A}(t))$$

(46)

where

$$\nu''^{-1} = \nu^{-1} + \nu'^{-1}$$

(47)

or in other words the effect of $\rho$ would appear to be to redefine the effective population size as

$$\nu \rightarrow \left( \frac{1}{\rho} + \frac{1}{\nu} \right)^{-1} n$$

(48)

As one would expect, for fixed $R_0$ and $k$ decreasing $\rho$ decreases the effective population size, since the variability due to sampling increases.
S8.1 Detailed analysis

While the simplified analysis above is already adequate for understanding the basic dynamics, it is necessary to do a more careful analysis to reveal some of the additional subtleties involved. In particular let us be careful about distinguishing between the unobserved allele frequencies \( \tilde{x}_a(t) \) (corresponding to all infected individuals) and the observed allele frequencies \( x_a(t) \) (corresponding to infected individuals whose genomes are sequenced). We can write

\[
\begin{align*}
[\text{infection}] & \quad \tilde{x}(t+1) \sim \mathcal{N}(\tilde{x}(t) + \tilde{d}(t), \nu^{-1} \tilde{\Lambda}(t)) \\
[\text{sampling}] & \quad x(t) \sim \mathcal{N}(\tilde{x}(t), \nu^{-1} \tilde{\Lambda}(t)) \\
[\text{sampling}] & \quad x(t+1) \sim \mathcal{N}(\tilde{x}(t+1), \nu^{-1} \tilde{\Lambda}(t+1))
\end{align*}
\]

where \( \tilde{d}(t) \) and \( \tilde{\Lambda}(t) \) are defined in terms of \( \tilde{x}(t) \) and \( \tilde{\Lambda}(t+1) \) is defined in terms of \( \tilde{x}(t+1) \), and where we have assumed that \( \rho \) and thus \( \nu' \) is constant across time. Ultimately we are interested in how \( x(t+1) \) depends on \( x(t) \), since these are the quantities we observe. To make progress in analyzing Eqn. 49-51 we first make the replacements \( \tilde{d}(t) \rightarrow d(t), \tilde{\Lambda}(t) \rightarrow \Lambda(t) \) and \( \tilde{\Lambda}(t+1) \rightarrow \Lambda(t+1) \), which we expect to be approximately valid in the diffusion limit \( n \rightarrow \infty \). With this substitution we find that \( x(t+1) \) is distributed as

\[
x(t+1) \sim \mathcal{N}(x(t) + d(t), \nu^{-1} \Lambda(t))
\]

If we now further assume that the drift is sufficiently small so that \( \Lambda(t) \approx \Lambda(t+1) \) we find that

\[
x(t+1) \sim \mathcal{N}(x(t) + d(t), \nu''^{-1} \Lambda(t))
\]

where

\[
\nu'' = \left( \frac{1}{\nu_0} + \frac{1}{k} + \frac{2}{\rho} \right)^{-1}
\]

so that Eqn. 48 was almost correct, except that—since we need to ‘denoise’ twice, i.e. once for time step \( t \) and once for time step \( t+1 \)—the \( \rho^{-1} \) term appears with an additional factor of two. It is easy to check numerically that the approximation in Eqn. 53 is a reasonably good one provided that \( n \) is large, \( \rho \) is small, and the drift is subdominant.

Eqn. 53 is reassuring because it means that sampling induces noise that is congruent with the noise induced by the Negative Binomial dynamics that underlies the discrete time process in Sec. S7. In particular this means that we do not need to separately estimate e.g. \( k \) and \( \rho \), since it suffices to estimate the effective population size in Eqn. 74. This is a significant simplification and is a key component in the viability of our overall approach.

S9 Prior inclusion probability

An important hyperparameter in our modeling approach is the prior inclusion probability \( h \) in Eqn. 29. In an equivalent parameterization we can instead consider the quantity

\[
S \equiv \mathbb{E}[\sum a_i \gamma_{a_i}] = hA
\]

where \( S \) is the expected number of non-neutral alleles a priori and the expectation in Eqn. 55 is with respect to the prior in Eqn. 29.

We consider two approaches to setting \( h \) (and thus \( S \)). The first is to set \( h \) by hand utilizing any a priori knowledge we may have of the expected number of non-neutral alleles. While a priori knowledge of \( h \) may be imprecise, this need not be particularly troubling: as we show in simulations in the main text results are not expected to be overly sensitive to the precise value of \( h \). Indeed, it is worth emphasizing that \( h \) is the prior inclusion probability. Sufficient evidence in the likelihood can always overwhelm the prior and the posterior inclusion probability (the PIP) can differ significantly from \( h \).

The second approach is to place a prior on \( h \) (and thus implicitly on \( S \)). A natural choice is to place a Beta prior on \( h \)

\[
h \sim \text{Beta}(\alpha_h, \beta_h)
\]

By Eqn. 50-51 \( x(t) \rightarrow \tilde{x}(t) \) and \( x(t+1) \rightarrow \tilde{x}(t+1) \) as \( n \rightarrow \infty \).
where $\alpha_h > 0$ and $\beta_h > 0$ are hyperparameters that control the Beta prior. This prior assumption has the following properties. The prior mean of $h$ is given by

$$E[h] = \frac{\alpha_h}{\alpha_h + \beta_h}$$

(57)

For $\beta_h \gg \alpha_h$ and $\beta_h \gg 1$ this mean is approximately equal to $\alpha_h/\beta_h$ and we further have that the variance of $h$ is approximately given by

$$\nabla[h] \approx \frac{\alpha_h}{\beta_h^2}$$

(58)

Consequently in this regime the ratio of the square root variance to the mean—which is a useful measure of the width of the prior distribution—is approximately given by the formula

$$\frac{(\nabla[h])^{\frac{1}{2}}}{E[h]} \approx \frac{1}{\sqrt{\alpha_h}}$$

(59)

These formulae are useful for choosing $\alpha_h$ and $\beta_h$. For example, if we have $A = 1000$ alleles and we expect that $S$ is about 10 and likely between 3 and 20 we might choose $\alpha_h = 1$ and $\beta_h = 100$, which concentrates about 90% of the prior probability mass between $S = 0.5$ and $S = 30$. Once we have chosen $\alpha_h$ and $\beta_h$ we must still infer $h$. In the absence of tempering this is easy to do since the posterior conditional distribution over $h$ is just another Beta distribution due to conjugacy. Unfortunately, the tempering introduced by the MCMC scheme we use spoils this relationship [4]. Nevertheless this conjugacy can easily be restored by introducing an additional auxiliary latent state in the MCMC algorithm as explained in detail in [5]. Thus in practice inferring $h$ is straightforward.

S10 Alternative modeling and inference strategies

In Sec. S1 we described how a diffusion-based likelihood can be used to model genomic surveillance data. In Sec. S2 we described the simplest inference strategy that utilizes such a likelihood, namely MAP inference. In Sec. S3 we described our own diffusion-based approach: Bayesian Viral Allele Selection. This method differs from MAP in two notable ways: i) it uses a different prior; and ii) it uses a different inference algorithm (MCMC instead of maximum a posteriori inference). In this section we describe a few alternative diffusion-based modeling and inference strategies.

S10.1 Laplace prior

We first describe the simplest modification of MAP that can account for the expected sparsity of non-neutral alleles. In this approach we place a Laplace prior on $\beta$

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

(60)

where $||\beta||_1$ is the $L^1$ norm of $\beta$ 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 [7]

$$\hat{\beta}^{\text{Laplace}} = \arg\max_{\beta} p(\beta|\sigma^{\text{Laplace}})p(y_{1:T-1}|\beta, \nu)$$

(61)

In order to compute $\hat{\beta}^{\text{Laplace}}$ we need to rely on optimization methods. In our experiments we use the Adam optimizer [6] and implement inference using the Pyro probabilistic programming framework [7]. In particular we use the Adam optimizer with an initial learning rate of $10^{-2}$ that exponentially decays to $10^{-4}$ over the course of $10^4$ optimization steps. Solving this optimization problem is

---

6To lighten the notation we assume a single spatial region.

7For the sake of precision we should probably refer to the approach in Sec. S2 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. One crucial difference between these two methods is that the former method admits a closed form solution, whereas the latter method requires iterative optimization techniques.

8The hyperparameters that control the momentum are $(\beta_1, \beta_2) = (0.5, 0.99)$, which is appropriate for this non-stochastic regime.
Another possible approach would be to consider the exact same model as in Eqn. 61 but to perform variational inference, for example mean-field variational inference utilizing a product of Normal distributions as the variational distribution over $\beta$. We would expect this approach to suffer from one of the main issues that plagues the approach in Eqn. 61: the hyperparameter $\sigma$ is difficult to set, but a well-chosen value of $\sigma$ is crucial for good performance.

S10.2 Horseshoe prior

Another approach to inducing shrinkage in $\beta$ would be to use a hierarchical shrinkage prior like the Horseshoe [8]. However, in our experience MCMC algorithms for Horseshoe-based regression can mix quite poorly in the high-dimensional setting. In addition, since the Horseshoe prior is a continuous shrinkage prior, it does not provide an interpretable allele-level score like the PIP provided by a Bayesian Variable Selection approach (see Eqn. 24). Moreover, MCMC inference in this context is likely to be expensive, since $A$ does not have low-rank structure that can be exploited for computational speed-ups.

S10.3 Discussion

Having now discussed a few possible alternatives to BVAS, some of the favorable characteristics of BVAS come into sharper focus. First, a fully Bayesian approach that systematically integrates over different hypotheses about which alleles are neutral and which are not contributes to reduced sensitivity to prior hyperparameters. Second, the sparsity assumption has notable computational benefits with the consequence that BVAS admits efficient MCMC inference with $O(\gamma^2 A)$ cost per iteration (e.g. compare to the quadratic dependence on $A$ in the Laplace method). Third, BVAS offers interpretable posterior inclusion probabilities. Fourth, the MCMC inference algorithm itself does not include any difficult-to-set hyperparameters, in contrast for example to the Laplace method above, which may require tweaking optimization hyperparameters for optimal performance.

S11 Note on Statistical Analysis

All statistical analysis in this work is done using standard Bayesian methods. In particular unless noted otherwise we report 95% credible intervals based on the (approximate) posterior distribution. Since all our models jointly consider all data—as opposed to univariate association tests—no corrections for multiple hypothesis testing are necessary.

S12 Limitations

We provide an extended discussion of some of the limitations of BVAS not touched upon in the main text. Due to limitations of our pre-processing pipeline that we inherit from UShER, we do not consider insertions and deletions. If there are insertions or deletions that have non-negligible selection effects, the exclusion of these alleles from our analysis may lead to inferences that incorrectly implicate other alleles that are linked to these insertions/deletions. These considerations are also applicable to our epistasis analysis, which may be particularly sensitive to any such missing allelic information. More broadly, the fidelity of the epistasis analysis may also be impacted by the exclusion of interaction effects outside the RBD.

Our method cannot correct for any issues in the raw input data (e.g. sequencing errors) and cannot correct for lineage-dependent sampling bias (e.g. sequencing only in case of S-gene target failure). Likewise biased inferences can result if, e.g., disease severity varies across variants and disease severity influences the likelihood that an individual’s viral genome is sequenced. Our diffusion-based likelihood cannot account for differential fitness effects that depend on unobserved confounders like population age structure. In addition we assume that individuals are infected with a single SARS-CoV-2 variant, an assumption that can be violated [9].

Additional bias and/or reduced statistical power may result from our use of a global effective population size, which implicitly assumes that the sampling rate does not vary wildly from region to region. In particular by (very likely) underestimating the effective population size in well-sampled regions like the UK we lose statistical power, although we note that by using a smaller effective
population size we expect our inferences to be less sensitive to noise in the data and thus more robust. Moreover by using a single effective population size we implicitly assume that the sampling rate is approximately uniform within each region. We do not expect a non-uniform sampling rate within a given region to bias results provided that the region is well mixed. Biased inferences may result, however, if a region is poorly mixed and exhibits non-uniform sampling, as this combination of violations of our modeling assumptions may overemphasize the importance of local fluctuations in allele frequencies.

S13  Simulations

S13.1  Simulation details (extended)

Unless otherwise noted, the parameters and distributional assumptions of all simulation-based experiments are as follows. We consider \( V = 100 \) viral variants and a true selection coefficient vector \( \beta^* \) with 10 non-zero coefficients:

\[
\beta^* = (0.01, 0.02, 0.04, 0.06, 0.08, -0.01, -0.02, -0.04, -0.06, -0.08, -0.10, 0, 0, ..., 0) \quad (62)
\]

In other words there are exactly 10 non-neutral alleles, 5 of which have positive effects on viral fitness and 5 of which have negative effects on viral fitness. We sample each element of each genotype vector \( g_v \) i.i.d. from a Bernoulli distribution with mean \( \frac{1}{V} \), i.e. \( g_{v,a} \sim \text{Bernoulli}(\frac{1}{V}) \) for \( v = 1, ..., V \) and \( a = 1, ..., A \) where \( A \) is the total number of alleles. This choice implies that typical variants have about two non-neutral alleles. We set \( R_0 = 1 \) so that typical reproduction numbers for variants \( v \) range between \( R_v \sim 0.9 \) and \( R_v \sim 1.1 \). In each simulation we consider a given number of \( R \) regions and \( T = 26 \) time steps. This latter choice corresponds to about one year of data assuming time bins of 14 days.

The initial number of infected individuals at time \( t = 1 \) within each region is drawn from a Negative Binomial distribution with mean \( 10^4 \) and dispersion parameter \( 10 \). Within each region these infected individuals are then apportioned to each of the \( V \) variants using a Multinomial distribution with uniform probabilities \( V^{-1} \). Thus the mean case count of each variant \( v \) within each region at time \( t = 1 \) is given by \( 10^4/V = 100 \). Case counts for \( t = 2, ..., T \) are then determined by the stochastic dynamics in Eqn. 1 with \( k = 0.1 \). This value of \( k \) is chosen since it consistent with estimates of the SARS-CoV-2 dispersion parameter \([10, 11, 12, 13]\). These raw counts are then subjected to Binomial sampling with mean \( \rho = 0.01 \), representing a sampling rate of 1\%, i.e. the viral sequences of 99\% of cases are not observed. Note that this implies that the expected number of observed individuals with variant \( v \) within each region at the initial time step is 1. Thus our parameter choices result in simulated data that is highly stochastic and that constitute a regime in which we expect that recovering the true selection coefficients \( \beta^* \) 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. Indeed from Eqn. [54] we would expect that

\[
\hat{\nu} \approx \left( \frac{1}{T} + \frac{1}{T} + \frac{2}{0.01} \right)^{-1} \times 10^4 \approx 0.01 \times 10^4 \approx 50 \quad (63)
\]

which is indeed what we observe, see Figure A. Note the fact that most regions have an effective population size somewhat larger than 50 is due to the fact that in most regions the number of infected individuals increases throughout the course of the pandemic.

S13.2  \( \hat{\nu} \) estimator performance

We investigate the accuracy of the effective population size estimators described in Sec. S6. Our simulated data are as in Sec. S13.1 except that we limit the duration of the pandemic to \( T = 10 \) time steps to limit the exponential growth/decay within any given region so that the true population size is approximately constant across time within any given region. We run the simulator for \( N_R = 10^5 \) regions. We then use Eqn. [54] to compute the ‘true’ effective population size for each region, where we average Eqn. [54] across \( T \) time steps (since the true number of infected individuals fluctuates

\^9The variance is thus \( \sim 3164^2 \).
Figure A: We depict the distribution of estimated effective population sizes $\hat{\nu}$ for the simulated data described in Sec. S13.1. The mean and standard deviation are 75.6 and 28.1, respectively. Note that the underlying true population size is about $10^4$.

Figure B: We evaluate the performance of our basic $\hat{\nu}$ estimator, Eqn. [36] to estimate the effective population size for each region.

In Figure B we depict histograms of ratios of the estimated $\hat{\nu}$ to the true effective population size $\nu_{true}$ for two scenarios. In the first scenario (left) we simulate data with $\beta = \beta^*$ as in Eqn. [62] with the result that drift is relatively moderate. In the second scenario (right) we simulate data with $\beta = 5\beta^*$, with the result that drift is much more pronounced. We find reasonably good agreement with a relative accuracy of about 10-40% depending on the scenario.

Figure B: We evaluate the performance of our basic $\hat{\nu}$ estimator, Eqn. [36]. We compare a scenario with smaller drift (left) and larger drift (right). See Sec. S13.2 for details. The ratio is approximately normally distributed with a mean of 1.02 and standard deviation of 0.09 (left) and a mean of 0.77 and standard deviation of 0.17 (right). Typical reproduction numbers $R_v$ range from about 0.9 to 1.1 (left) and about 0.5 to 1.5 (right).

S13.3 Method Comparison

We provide some additional information regarding the method comparison in the main text. We use the following hyperparameter settings. For MAP we test all values of $\tau$ ranging from $2^{-24}$ to $2^{24}$ in powers of 2 and choose $\tau = 2^{11} = 2048$, since this value gives the best results across the board. In other words we are using the ground truth to choose $\tau$, something that we obviously cannot do when analyzing real data. Since the effective regularization parameter for MAP is given by $\gamma_{reg} = \tau/\nu$ (see Eqns. [22]) and $\nu \approx 75$ for our simulated data this corresponds to a value of $\gamma \approx 30$, which is in the range of values suggested by [1]. For BVAS we use $\tau = 100$ and $(\nu_h, \beta_h) = (1/16, A/32)$ except for experiments where we explicitly vary the prior inclusion probability $h$. Unless specified otherwise these BVAS settings are used in all simulation experiments. For Laplace we use a regularization scale of $\sigma_{Laplace} = 0.01$. For PyR0 we use the default settings in [2]. Unless specified otherwise, we

10We found that it is only for $A = 100$ where there is a strong preference for a smaller value of $\tau$ like $\tau = 2^9$; for larger $A$ there is a uniform preference for larger values of $\tau$.

11This corresponds to approximately 2 expected non-neutral alleles a priori and encodes substantial a priori uncertainty about $h$. See Sec. S9 for discussion. Note that we use a prior over $h$ instead of specifying $h$ as a fixed hyperparameter to avoid leveraging the fact that we know that the true number of non-neutral alleles is given by 10.

13
use the global $\hat{\nu}$ estimator described in Sec. \ref{sec:global-nu}. We note that we tried using z-scores to rank PyR$_0$ hits but found essentially identical results. We also note that the hit rate can be understood as the precision that results if the top 10 alleles are declared hits. We run BVAS for 5500 iterations, discarding the first 500 samples for burn-in. We use $T_{\text{opt}} = 10^4$ optimization steps when running Laplace.

Figure C: We depict the precision and sensitivity performance of BVAS using the simulated data described in Sec. \ref{sec:validation}. For this purpose we define hits as all alleles with PIPs above a threshold of 0.1. Precision is the number of true positives divided by the total number of hits. Sensitivity is the fraction of causal alleles that are identified as hits.

Figure D: We depict how the hit rate (left) and root mean squared error (RMSE; right) vary as a function of the sampling rate. We consider $A = 2000$ alleles and $N_R = 30$ regions.

Figure E: We explore BVAS performance in a scenario with vaccination-dependent selection coefficients. Precision computations assume a PIP cutoff of 0.1. Note that we only count a hit as correct if it is correctly identified as $\phi$-dependent or $\phi$-independent.

\subsection*{S13.4 Comparing global vs. regional $\hat{\nu}$ strategies}

In the main text we investigated the effect of modulating the overall scale of $\nu$. It is also important to investigate whether we should prefer the global or regional $\hat{\nu}$ estimators introduced in Sec. \ref{sec:global-nu}. As can be seen in Figure\ref{fig:vavsa}, the two strategies perform nearly identically, at least on our simulated data. This suggests that the more ‘ambitious’ regional estimator \emph{could} perform well in real data, but it also suggests that it may be entirely sufficient to use the (presumably) more robust global estimator. For this reason we only use the global estimator in our actual analysis. By virtue of this choice all regions in our analyses contribute equally to the likelihood.

\footnote{This is only relevant for the diffusion-based methods. The particular value of $\nu$ used is most relevant for BVAS and Laplace and is somewhat less important for MAP, which only depends on $\nu$ through the quantity $\gamma_{\text{reg}} = \tau/\nu$.}
S13.5 Spike-in experiment

SARS-CoV-2 data are prepared as in Sec. S14.1. Then for each variant v in the data the genotype g_v is extended by 200 fake spike-in alleles, where we sample the genotype of each spike-in allele i.i.d. from a Bernoulli distribution with mean 0.01.

S13.6 Including vaccination-dependent effects

Our simulation is as in Sec. S13.1 except we now assume 20 non-zero effects, 10 of which are vaccination-dependent and 10 of which are not (we assume these two sets of ten are disjoint and that each set has the same effect sizes as the non-zero effects in Eqn. 62). We assume that ϕ_r(t) starts at zero everywhere (i.e. ϕ_r(t = 1) = 0), increases linearly over time, and that the final value of ϕ_r(t) in each region is given by ϕ_r(T) ∼ Uniform(1/2, 1).

S14 SARS-CoV-2 Analysis

S14.1 Data Pre-processing

The initial steps of our data pre-processing pipeline follow the procedures described in [2] and make use of the open source code at https://github.com/broadinstitute/pyro-cov. We give a short summary of the relevant parts of this pipeline and refer the reader to [2] for additional details.

Our initial data ingest consists of 8,575,902 samples downloaded from GISAID [14] on April 18th, 2022. Each sample record includes metadata for time of collection, geographic location, and PANGO lineage annotation [15] as well as genetic sequence data. We discard records without complete metadata. Sequences whose alignment quality is not reported as “good” are discarded. Time intervals are binned into 14-day segments, which results in a total of 62 time intervals, i.e. a bit more than two years of data. Note that we choose a multiple of 7 days to minimize weekly seasonality effects (e.g. closure of testing facilities on weekends).

S14.1.1 Lineage Clustering

Since the likelihood that underlies BVAS is defined in allele frequency space, it does not explicitly rely on notions of variant or lineage. However, the pre-processing pipeline in [2] makes use of a dynamic partitioning of genetic samples into clusters. One potential advantage of this approach is that it can remove sequencing artifacts. Like [2] our default analysis makes use of L = 3000 clusters derived from the 1694 PANGO lineages present in our dataset. That PANGO lineages are not sufficiently fine can be seen in Figure G, which shows that some PANGO lineages (notably B.1.1) consist of several sublineages with widely varying fitnesses. This recapitulates the findings in [2]. These L = 3000 clusters are created as follows. The starting point is a 9,293,486 node phylogeny of all GISAID samples maintained by Angie Hinrichs [16]. This phylogeny was created using UShER [17] and excludes private mutations, masks difficult-to-sequence regions, disregards deletions, and parsimoniously imputes missing sequence data. To coarse grain the 9.3 million node phylogenetic tree down to L = 3000 clusters we iteratively collapsed parent-child edges by greedily minimizing a notion of distance defined between pairs of mutation-annotated trees. Empirically [2] found that this heuristic clustering produces trees that are approximately balanced (both in terms of cluster size and cluster-cluster edit distance). Note that this clustering procedure removes some of the genetic
diversity in the raw data. The total number of alleles retained at the end of clustering is $A = 2975$. In Figure 1 we investigate the sensitivity of our analysis to the choice of $L$ and find only modest differences.

S14.1.2 Spatial Aggregation

To accommodate the fact some spatial regions are sparsely sampled while other spatial regions are densely sampled, we dynamically aggregate regions to encourage approximately balanced sample totals per region. For example, in well-sampled countries like the US and Germany our spatial regions are at the state/province level (e.g. California and Bavaria) while subregions of less densely sampled countries are aggregated up to the country level (e.g. Turkey, Portugal, and Slovenia).

The result of the above pre-processing steps is a collection of region-specific time series of lineage counts (encoded as a $R \times T \times L$ count array) as well as lineage-level amino acid features (encoded as a $L \times A$ binary array with $A = 2975$).

S14.1.3 Allele Frequency Space

Running BVAS MCMC requires two basic ingredients (see Eqn. 28): $\tilde{y}^\nu$ and $\tilde{\Lambda}^\nu$. Since these allele frequency space quantities are unique to BVAS and the method in [1] the rest of our pre-processing pipeline diverges from [2].

S14.1.4 Diffusion Limit Filters

BVAS relies on the diffusion limit, i.e. it assumes that the number of infected individuals is relatively large. For this reason we would like to exclude regions that are poorly sampled from our analysis. With that caveat in mind we would like to include as many regions as possible so that the results of our analysis are not unduly influenced by the idiosyncracies of any particular region. To manage this trade-off we introduce two hyperparameters: $N^\nu_{\text{min}}$ and $N^{14\nu}_{\text{min}}$. Regions where the total number of samples is less than $N^\nu_{\text{min}}$ are not included in our analysis. By making $N^\nu_{\text{min}}$ relatively large we exclude regions that have only been observed for a handful of time intervals (note that it is particularly difficult to estimate the effective population size of such regions). We must also decide which time intervals to include in our analysis. Even if the total number of samples in a given region is large the number of samples collected early on in the pandemic may be small, with the result that dynamics in that early phase of the pandemic will tend to be dominated by noise (i.e. diffusion). For each region we thus only include those 14-day time intervals for which the number of samples is greater than or equal to $N^{14\nu}_{\text{min}}$. Moreover, since our likelihood depends on allele frequency increments $y(t)$ (see Eqn. 12) we choose the longest subsequence of contiguous time intervals where each time interval contains at least $N^\nu_{\text{min}}$ samples. In our default analysis, we choose relatively conservative values of these two hyperparameters: $N^\nu_{\text{min}} = 5 \times 10^3$ and $N^{14\nu}_{\text{min}} = 50$. In Sec. S14.2 we investigate the sensitivity of our estimates to $N^\nu_{\text{min}}$ and $N^{14\nu}_{\text{min}}$. In Table B we list the 128 regions that enter into our default analysis with $N^\nu_{\text{min}} = 10^3$ as well as the 67 regions that enter into an auxiliary analysis with $N^{14\nu}_{\text{min}} = 20 \times 10^3$.

S14.1.5 Computing $\tilde{y}^\nu$ and $\tilde{\Lambda}^\nu$

The next step is to compute $\tilde{y}^\nu$ and $\tilde{\Lambda}^\nu$. First we estimate the effective population size $\hat{\nu}$ for each region using the estimator in Eqn. 36. In our default global analysis we then compute the median $\nu$ across all the regions included in the analysis (note that the number of regions included depends on $N^\nu_{\text{min}}$) and scale the median value of $\hat{\nu}$ by a factor of 0.5. This choice is motivated by several concerns. The median effective population size $\hat{\nu}$ in our default analysis with 128 regions is 63.9. This value exceeds the estimated effective population size of 64 of the 128 regions, see Table B Moreover, half that value, namely 32.0, exceeds the estimated effective population size of only 9 of the 128 regions. Thus by making this choice we ensure that $\hat{\nu}$ is not grossly overestimated in any one region (the smallest estimated $\hat{\nu}$ is 20.0 in Fukuoka, Japan). Consequently, with this choice BVAS is not expected to confuse the inherent stochasticity of the least well sampled regions with signal.

In any case we do not expect to learn very much from regions with small effective population sizes, since the dynamics in these regions is dominated by noise. In this context see Figure S23 in [2], which shows that PyR0 results are largely driven by the most well-sampled regions.

13In any case we do not expect to learn very much from regions with small effective population sizes, since the dynamics in these regions is dominated by noise. In this context see Figure S23 in [2], which shows that PyR0 results are largely driven by the most well-sampled regions.

14This is especially important for accommodating vaccination-dependent effects as in Eqn. 42.
To put this choice in a different light, we note that Bayesian inference has a long history of grappling with model mis-specification and robustness. One class of methods achieves robustness by raising the likelihood to a power, which is precisely the result of scaling the effective population size. Thus our choice of $\hat{\nu}$ can be understood as a way of achieving robustness against possible model mis-specification. In particular we do not expect our genomic surveillance data to exactly follow the discrete time branching process in Sec. S1 nor do we expect the reproduction number to be exactly linear in the genotype. For more discussion on these and related ideas see [18].

With this choice of $\hat{\nu}$ we now compute $\hat{y}^{\nu}$ and $\hat{\Lambda}^{\nu}$ using the definitions in Eqn. 31 and using a shared value of the effective population size across all regions. Note that computing these quantities is $\mathcal{O}(NRT)$, i.e. the computation scales with both the number of regions and the number of time intervals under consideration, but that $\hat{y}^{\nu}$ and $\hat{\Lambda}^{\nu}$ can be computed very efficiently in an online fashion. Indeed computing $\hat{y}^{\nu}$ and $\hat{\Lambda}^{\nu}$ only takes a few seconds on a mid-grade GPU. Importantly, these $\mathcal{O}(NRT)$ computations only need to be done once in pre-processing, since the BVAS MCMC algorithm operates directly on $\hat{y}^{\nu}$ and $\hat{\Lambda}^{\nu}$.

We also describe an alternative strategy for computing the effective population size that we use in our analysis of SARS-CoV-2 surveillance data obtained through August 10th 2022. As above we compute effective population size estimates for each region using Eqn. 36. We then partition the regions into two buckets based on the median effective population size: i) those whose $\hat{\nu}$ exceeds the median; and ii) those whose $\hat{\nu}$ is less than or equal to the median. That is we partition regions into lower and upper 50% quantiles based on their estimated effective population sizes. We then compute $\hat{y}^{\nu}$ and $\hat{\Lambda}^{\nu}$ using the median $\hat{\nu}$ for the upper bucket and the minimum $\hat{\nu}$ (i.e. the minimum among all regions) for the lower bucket. We call this effective population size estimation strategy the ‘two-bucket’ strategy. It has the following nice properties: i) it is parameter free; ii) it weighs data from well-sampled regions like the UK more than poorly-sampled regions (though only by a moderate $\mathcal{O}(1)$ factor); and iii) it is still conservative in that we generally expect to underestimate the true effective population size in each region.

S14.1.6 Vaccination rates

Vaccination data are downloaded from OWID [19]:

- https://github.com/owid/covid-19-data/raw/master/public/data/vaccinations/vaccinations.csv
- https://github.com/owid/covid-19-data/raw/master/public/data/vaccinations/us_state_vaccinations.csv

We use two definitions of vaccination status: people_vaccinated_per_hundred and people_fully_vaccinated_per_hundred. We use country-level data for all regions except for US states and England, Scotland, Wales. Any missing values are interpolated using linear interpolation. We consider the same 128 regions as in our default analysis, except we exclude Luxembourg, since it does not include complete vaccination information. Thus we consider 127 regions in total.

S14.1.7 MCMC

When running BVAS we do $51 \times 10^4$ MCMC iterations. We discard the first $10^4$ samples and retain the remaining $5 \times 10^5$ samples to compute posterior quantities of interest like PIPs. All computations are done in 64-bit precision.

S14.2 Sensitivity Analysis

See Figure M-AX and the main text for discussion.

S14.3 Backtesting

We follow the same data pre-processing pipeline as in Sec. S14.1 although we note that for any given date fewer than 128 regions enter our analysis, since we are considering subsets of the data and the $\Lambda_{\min}^{\text{tot}} = 10^4$ requirement is still in place. In practice a real-time surveillance program might choose to use looser filters (i.e. smaller $N_{\min}^{\text{tot}}$ and $N_{\min}^{14}$).

\[15\] In particular we avoid instantiating large arrays whenever possible. For example, we do not compute $\hat{\Lambda}^{\nu}$ separately for each region and then sum over regions in an outer loop; rather we accumulate a single $A \times A$ matrix as we go.
Table A: We report the relative growth rates $R/R_A$ of recombinant SARS-CoV-2 lineages as estimated by BVAS. Uncertainty estimates denote 95% credible intervals.

| Growth Rate | Growth Rate |
|-------------|-------------|
| XN 7.329 ± 0.689 | XM 6.934 ± 0.721 |
| XT 7.273 ± 0.706 | XK 6.736 ± 0.717 |
| XR 7.249 ± 0.715 | XS 6.290 ± 0.599 |
| XL 7.206 ± 0.722 | XP 6.216 ± 0.600 |
| XG 7.197 ± 0.722 | XD 5.848 ± 0.651 |
| XQ 7.147 ± 0.715 | XF 5.787 ± 0.575 |
| XH 7.012 ± 0.718 | XC 3.578 ± 0.487 |
|XE 6.963 ± 0.722 | XB 2.447 ± 0.406 |
|XJ 6.962 ± 0.722 |XA 2.268 ± 0.236 |

Figure G: We depict the 10 most diverse PANGO lineages as estimated by our model. The family of B.1.1 lineages is notably diverse, with some of the most recent sublineages—the dates of emergence for a few select lineages are annotated—exhibiting growth rates that exceed that of Delta and approach that of the least fit Omicron sublineages.

S14.4 Analysis using data through August 10th 2022

We report BVAS inference results obtained using SARS-CoV-2 data obtained through August 10th 2022. This analysis considers 151 regions and 84,977,079 viral sequences. To estimate effective population sizes we use the ‘two-bucket’ strategy described in Sec. S14.1. Otherwise our analysis uses the same hyperparameters as in our default analysis using data through April 18th, 2022. For results see Table C and Table D.

S14.5 Comparison to MAP

In Figure DD we compare BVAS results to results obtained with MAP. Estimates for selection coefficients $\beta$ are qualitatively similar to BVAS results, although BVAS results are much sparser, with many coefficients nearly zero (Figure DD). Growth rate estimates are in good agreement if the MAP regularizer $\tau$ is set sufficiently low ($\tau = 2^8$ or $\tau = 2^{10}$), but otherwise MAP estimates appear to be over-regularized towards Wuhan A, see Figure EE. Conversely, Manhattan plots for MAP estimates are dense unless $\tau \geq 2^{12}$ (Figure HH). Thus MAP is unable to simultaneously yield large growth estimates for fit lineages like Omicron and assign negligible effect sizes to most alleles. If we believe that most alleles should be neutral, this is a limitation of MAP. More generally, MAP offers limited control over how much estimates should be regularized because it is controlled by a single hyperparameter $\tau$. This is in contrast to BVAS where regularization is controlled by both $S$ and $\tau$. Indeed BVAS uses a value of $\tau_{\text{reg}} = \tau/\nu$ that is an order of magnitude smaller than MAP, depending on $S$ for additional regularization. The analysis in Figures EE, HH indicates that this additional regularization is crucial. We also note that MAP uncertainty estimates are quite
Figure H: For the 1729 SARS-CoV-2 clusters with at least one amino acid substitution in the RBD domain we compare: i) the BVAS prediction for the contribution to $\Delta R$ from RBD substitutions only; to ii) antibody binding computed using the antibody-escape calculator in [20]. The escape calculator is based on deep mutational scanning data for 33 neutralizing antibodies elicited by SARS-CoV-2. BVAS predictions exhibit high (Spearman) correlation with predictions from [20]. Compare to Figure 2B in [2].

Figure I: View of the 419 amino acids of the nucleocapsid (N) protein domains, annotated by structure [21]. Many top-scoring mutations occur in the serine–arginine rich region (SR), reported to be immunogenic by [22].

narrow (see e.g. Figure EE) which can ultimately be traced to the fact that MAP considers a single hypothesis about allele neutrality (namely that no alleles are neutral).

Note that our MAP analysis follows the exact same data pre-processing pipeline as for BVAS, see Sec. S14.1. Thus BVAS and MAP comparisons utilize the exact same likelihood, although we do not scale the $\hat{\nu}$ estimate in the case of MAP and instead use the median effective population size directly. Also note that in Figure GG and Figure II we choose $\tau = 2^{12}$, which corresponds to a level of regularization that appears to achieve the best balance between over- and under-regularization—although as we have argued above this compromise appears to be suboptimal.

S14.6 Comparison to PyR$_{0}$

We also compare BVAS results to results obtained with PyR$_{0}$, see Figure JJ-NN. While PyR$_{0}$ hits cluster in similar parts of the SARS-CoV-2 genome as BVAS hits (Figure JJ) and many of the top
Figure J: We depict Manhattan plots zoomed-in on two regions of the SARS-CoV-2 genome. (Top) View of the ORF1a polyprotein, including 11 non-structural proteins (nsps). (Bottom) View of the ORF1b polyprotein, including nsp12-16; note the amino acid positions do not account for 9 additional residues at the N-terminus of nsp12 (RNA polymerase) resulting from the -1 ribosomal frameshift.

hits are common to both methods (Figure KK), the quantitative agreement between PyR₀ and BVAS is modest. For example, only 7 hits are common to the top 25 BVAS and top 25 PyR₀ hits, and the Pearson correlation for β estimates is $R = 0.365$, see Figure NN. Moreover, PyR₀ estimates for the fittest Omicron lineages are substantially higher than BVAS estimates, by as much as ~ 50%, see Figure NN. Additionally PyR₀ estimates, especially for allele-level quantities like selection coefficients, are implausibly small, orders of magnitude smaller than the corresponding BVAS uncertainties.

We note that, like our Laplace method (see Eqn. 60), PyR₀ has a hyperparameter $\sigma_3$ that controls the sparsity of the selection coefficients. The analysis in [2] uses $\sigma_3 = 5 \times 10^{-4}$. However, our analysis includes approximately ~ 35% more data than the analysis in [2], suggesting that a smaller value of $\sigma_3$ may be required. Consequently, after comparing inference results for several values, we settle on $\sigma_3 = 2 \times 10^{-4}$, although we note that other similar values are also likely defensible. This highlights
one of the disadvantages of PyR₀: it depends on several hyperparameters that can be difficult to set in practice.

PyR₀ comparisons follow the data pre-processing pipeline described in [2]. We use the default parameters listed in [2], with the difference that we decrease σ₃ (as discussed above). Note that unlike BVAS the PyR₀ analysis does not make use of filters like N_{tot min}^4, and thus considers more data from significantly more regions (1694 instead of 128).

S14.7 Hitchhiking

As is well known, hitchhiking, i.e. apparent drift due to genetic linkage, complicates the inference of selection effects and makes it difficult to unambiguously disentangle driver and passenger mutations. Hitchhiking can involve both selective sweeps caused by the uptake of beneficial mutations [23] as well as the elimination of neutral variants linked to deleterious mutations, i.e. background selection [24]. One of the advantages of the diffusion-based likelihood in Eqn. 12 is that covariance information from Λ(t) can help distinguish driver mutations from passenger mutations. To make this more explicit, we identify the 3 alleles that exhibit large uptakes in allele fractions while simultaneously exhibiting small PIPs below 0.01: N:R203K, ORF14:G50N, and N:G204R. As can be seen in Figure 10, these alleles have in fact peaked three times, having become very prevalent during the B.1.1, B.1.177 (Alpha), and BA.1 (Omicron) waves, while experiencing significant drops in interim periods when other lineages (e.g. B.1.177 and Delta) predominated. BVAS infers that these three alleles are passenger mutations whose rising and falling prevalences are best explained by selection effects due to other linked alleles. While it is difficult to validate whether this particular inference is correct, it is reassuring that all three alleles have similar PIPs, as they exhibit near perfect co-occurrence in the data.
Figure L: The distribution of estimated effective population sizes $\hat{\nu}$ for the 128 regions in our default analysis. The smallest and largest values of $\hat{\nu}$ are 20.0 and 239.4, respectively. The vertical dotted lines correspond to the median effective population size (63.9; right) and half the median effective population size (32.0; left). The latter quantity is used in our default analysis (see Sec. S6 and Sec. S14.1).

Figure M: We depict the sensitivity of BVAS estimates of the top 10 lineages w.r.t. the hyperparameter $S$. The results of our default analysis are in blue. Here and elsewhere uncertainty estimates are 95% credible intervals.

S14.8 Vaccination-dependent effects

We follow the same data pre-processing pipeline described in Sec. S14.1 noting that the dimension of $y(t)$ doubles as we include vaccination-dependent effects. Note, however, that when we estimate the effective population size with Eqn. 36 we do so using allele frequency increments in the original $A$-dimensional space so that our $\hat{\nu}$ estimate remains the same.

S14.9 Epistasis

We report additional details pertaining to our epistasis analysis. We first note that the variant-space dynamics in Eqn. 3 are related to the allele-space dynamics in Eqn. 8 via a linear transformation specified by a genotype matrix. This still holds true when there are higher order interactions in the fitness ansatz in Eqn. 7. Consequently all the relevant formulae in Sec. S1 remain valid, with the difference that allele-space quantities for pairwise features must be re-interpreted appropriately. For example, entries $((a_1, a_2), (a_3, a_4))$ in the diffusion matrix $\Lambda(t)$ in Eqn. 10 corresponding to pairwise features $(a_1, a_2)$ and $(a_3, a_4)$ encode quartic moments w.r.t. the underlying amino-acid-level alleles.

Our epistasis analysis uses the same data and pre-processing as our default analysis. We use twice as many MCMC iterations as in our default analysis. For pairwise interaction effects we use an a priori inclusion probability $h_{\text{quad}}$ that is one half of the value used for linear effects, i.e. $h_{\text{quad}} = \frac{1}{2} h_{\text{linear}} \approx 0.0084$. We verified that other values of $h_{\text{quad}}$—in particular $h_{\text{quad}} = \frac{1}{4} h_{\text{linear}}$ and $h_{\text{quad}} = h_{\text{linear}}$—identified the same leading pairwise interaction effects (although with a corresponding shift in the overall PIP level for quadratic effects), with (N501Y, P681H) the leading hit in all three analyses. This remains true for sensitivity analyses in which $\hat{\nu}$ is modulated down by a factor of one half or the prior precision $\tau$ is halved.
Figure N: We depict the sensitivity of BVAS estimates of the top 10 lineages w.r.t. the number of regions in the analysis. The results of our default analysis are in blue. Note that the number of regions is effectively controlled by $N_{\text{min}}^{14}$, see Sec. S14.1.

Figure O: We depict the sensitivity of BVAS estimates of the top 10 lineages w.r.t. the hyperparameter $N_{\text{min}}^{14}$ (see Sec. S14.1). The results of our default analysis are in blue.

As touched upon in the main text, we only consider quadratic interactions between pairs $(a_1, a_2)$ of non-synonymous amino acid substitutions that co-occur in at least 2 of our $L = 3000$ SARS-CoV-2 clusters. Moreover for each $(a_1, a_2)$ we require that there are at least two clusters that ‘separate’ the pair, i.e. clusters that carry the $a_1$ allele but not the $a_2$ allele or vice versa. After applying these filters there remain 1432 pairwise interaction effects for consideration, where the vast majority of pairwise interactions that do not pass our filters are excluded because of the co-occurrence requirement. Note that BVAS would not have any incentive to make use of these pairwise interactions if they were included in the analysis, since they would be unidentifiable in the context of their constituent linear effects (and as such they do not add any additional expressivity). See Figure S15 for a sensitivity analysis for (linear) effects common to the epistasis analysis and the default analysis without epistasis. Top hits from our epistasis analysis are reported in Table F. The ‘pairwise’ posterior inclusion probabilities for these hits are depicted graphically in Figure S16.

In the main text we report that top-scoring interaction effects in our epistasis analysis are enriched for interactions between mutations that correspond to top-scoring (linear) effects in our default analysis. In particular 8 of the 10 amino acids in the top 8 quadratic hits in our epistasis analysis exhibit PIPs larger than 0.3 in our default analysis. Only A67V and E484A exhibit negligible PIPs in our default analysis but participate in a top 8 quadratic hit, namely the interaction between A67V and E484A.

The enrichment analysis is done as follows. For the 421 amino acid mutations $a$ in Spike under consideration, we assign a score given by $\sigma(a) = 1/\text{rank}$, where rank is the PIP rank of the corresponding linear effect in our default analysis (i.e. $\sigma(a) \in \{1/1, 1/2, 1/3, \ldots\}$). We then assign interaction effects between amino acids $a_1$ and $a_2$ a score given by the product $\sigma(a_1)\sigma(a_2)$. That is we assign high scores to interaction effects whose corresponding amino acids exhibit large PIPs in our default analysis. We then compute a Spearman correlation coefficient between the PIP ranks for interaction terms in our epistasis analysis with the scores $\sigma(a_1)\sigma(a_2)$. We find a moderate positive correlation given by 0.45. This substantiates our claim that top-scoring interaction effects in our epistasis analysis are enriched for interactions between mutations that correspond to top-scoring (linear) effects in our default analysis. To determine the statistical significance of this correlation we conduct a permutation test with 9999 permutations, resulting in a p-value less than 0.0001, i.e. none
Figure P: We depict the sensitivity of BVAS estimates of the top 10 lineages w.r.t. the hyperparameter $\tau$, which controls the prior variance of selection coefficients $\beta$. The results of our default analysis are in blue.

Figure Q: We depict the sensitivity of BVAS estimates of the top 10 lineages w.r.t. three values of effective population size estimate $\hat{\nu}$. The results of our default analysis are in blue. See Sec. S14.1 for details on how we choose $\hat{\nu}$.

S15 Bayesian variable selection

We briefly review the literature on Bayesian variable selection (BVS). Some of the earliest approaches to BVS were introduced by \cite{25,26}. \cite{3} provide an in-depth discussion of BVS for linear regression and CART models. \cite{4} introduce Tempered Gibbs Sampling (TGS) and apply it to BVS for linear regression. \cite{27} and \cite{28} review various methods for BVS.

\footnote{If we instead use an additive score $\sigma(a_1) + \sigma(a_2)$ we find a correlation of 0.19 and a p-value of 0.0058 \approx 0.006.}
Figure R: We depict the concordance of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for two analyses run with independent MCMC chains. The near perfect concordance between the two chains suggests that MCMC error is minimal.

Figure S: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for an analysis with the default number of regions ($N_R = 128$) and $N_R = 67$. Note that the number of regions is effectively controlled by $N_{\text{tot}}^\text{min}$, see Sec. S14.1.

Figure T: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for an analysis with the default number of SARS-CoV-2 fine lineage clusters ($L = 3000$) versus $L = 5000$ (see Sec. S14.1). The latter analysis includes $A = 3790$ alleles; here we subset to the $A = 2975$ alleles that are common to both analyses. The top scoring hit that appears in the $A = 3790$ analysis but does not appear in the $A = 2975$ analysis is ORF1a:K1869R with a PIP of 0.068 and $\beta = 0.013$. 

25
Table B: Summary statistics for the 128 regions used in our main analyses sorted by raw count totals. All regions are used in our default analysis with $N_{\text{tot}}^\min = 10^{10}$ and the top 67 regions are used in our secondary analysis with $N_{\text{tot}}^\min = 20 \times 10^{10}$. The 2nd and 3rd columns indicate the total number of samples in each region before and after filtering. The final column denotes the estimated effective population size $\hat{\nu}$. The total number of samples in all regions is 7,127,962 (raw) and 6,910,018 (filtered). The total number of samples in the top 67 regions is 6,246,448 (raw) and 6,106,650 (filtered).

Approximately 3% of the samples are removed in filtering due to the $N_{\text{tot}}^\min$ filter (see Sec. S14.1). The average $\hat{\nu}$ in all 128 regions is 77.1; in the top 67 regions it is 92.8. Note that some countries like Germany include both province-level spatial units like Bavaria and larger units that are aggregated from several provinces that are more sparsely sampled (see row 70).
Figure U: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for an analysis with the default prior precision ($\tau = 100$) versus $\tau = 25$.

Figure V: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for an analysis with the default prior precision ($\tau = 100$) versus $\tau = 400$.

Figure W: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/RA$ (right) for an analysis with the default expected number of non-neutral alleles a priori ($S = 50$) versus $S = 25$. The consequence of halving $S$ is that the total PIP sum decreases from 115.4 to 84.2 and the sum of all PIPs that are larger than 0.1 decreases from 64.2 to 56.4.
Figure X: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/R_A$ (right) for an analysis with the default expected number of non-neutral alleles a priori ($S = 50$) versus $S = 100$. The consequence of doubling $S$ is that the total PIP sum increases from 115.4 to 171.4 and the sum of all PIPs that are larger than 0.1 increases from 64.2 to 78.0.

Figure Y: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/R_A$ (right) for an analysis where the effective population size estimate $\hat{\nu}$ is 16.0 instead of the default 32.0. Notably, there are fewer sizable PIPs in the analysis with smaller $\hat{\nu}$. See Sec. S14.1 for details on how we choose $\hat{\nu}$.

Figure Z: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/R_A$ (right) for an analysis where the effective population size estimate $\hat{\nu}$ is 63.9 instead of the default 32.0. Notably, there are more sizable PIPs in the analysis with larger $\hat{\nu}$. See Sec. S14.1 for details on how we choose $\hat{\nu}$.
Figure AA: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/R_A$ (right) for an analysis where the hyperparameter $N_{14}^{\text{min}}$ is changed from the default of $N_{14}^{\text{min}} = 50$ to $N_{14}^{\text{min}} = 100$.

Figure BB: We investigate the ability of BVAS to estimate the fitness of lineages as they emerge. We consider several Delta (top two rows) and Omicron (bottom two rows) sublineages as they emerged in Spring 2021 and the end of 2021, respectively. In both cases we compare to the fittest lineage that was predominant (in England and elsewhere) before the rise of Delta and Omicron. We also depict time series of the number of SARS-CoV-2 sequences in our dataset for each sublineage.
Figure CC: In this companion figure to Figure BB we depict the stability of growth rate estimates for 9 ‘well-established’ lineages as large numbers of SARS-CoV-2 sequences were collected during the Omicron wave. In particular, the total number of sequences included in our analysis increased by about 25%, from 3,901,038 sequences on November 10th to 4,889,119 sequences on January 12th.

Table C: The 20 SARS-CoV-2 lineages with the highest (relative) growth rates $R_v/R_A$ as estimated by BVAS for data obtained through August 10th 2022.
Table D: The top 25 fitness-associated SARS-CoV-2 mutations as estimated by BVAS and ranked by PIP (left) and $\beta$ (right) using data obtained through August 10th 2022. Uncertainty estimates are 95% credible intervals.

| Mutation     | RBD | PIP | Beta |
|--------------|-----|-----|------|
| 1            | S.L452R | Yes | 1.0000 | 0.899 ± 0.140 |
| 2            | S.L452Q | Yes | 1.0000 | 0.490 ± 0.087 |
| 3            | S.P454K  | Yes | 1.0000 | 0.548 ± 0.045 |
| 4            | S.E484K  | Yes | 1.0000 | 0.396 ± 0.087 |
| 5            | S.R346T  | Yes | 1.0000 | 0.541 ± 0.119 |
| 6            | S.L452M  | Yes | 1.0000 | 0.452 ± 0.119 |
| 7            | S.L452M  | Yes | 1.0000 | 0.454 ± 0.105 |
| 8            | S.K468K  | Yes | 1.0000 | 0.411 ± 0.055 |
| 9            | S.S477N  | Yes | 1.0000 | 0.383 ± 0.113 |
| 10           | S.F490V  | Yes | 1.0000 | 0.523 ± 0.121 |
| 11           | ORF1ab.D10G | 1.0000 | 0.212 ± 0.073 |
| 12           | M.R72T   | Yes | 1.0000 | 0.346 ± 0.073 |
| 13           | S.P151H   | Yes | 1.0000 | 0.305 ± 0.104 |
| 14           | S.I17K    | Yes | 1.0000 | 0.233 ± 0.083 |
| 15           | M.D3N    | Yes | 1.0000 | 0.322 ± 0.073 |
| 16           | S.N440K   | Yes | 1.0000 | 0.314 ± 0.121 |
| 17           | S.K477N   | Yes | 0.9999 | 0.395 ± 0.109 |
| 18           | ORF1b.S666F | 0.9999 | -0.535 ± 0.075 |
| 19           | ORF1a.A0520V | 0.9995 | 0.084 ± 0.035 |
| 20           | S.A022Y   | 0.9995 | 0.193 ± 0.073 |
| 21           | ORF1a.A0520V | 0.9994 | -0.123 ± 0.033 |
| 22           | ORF1b.S1390R | 0.9990 | -0.058 ± 0.042 |
| 23           | S.M679K   | 0.9986 | 0.386 ± 0.176 |
| 24           | ORF1a.G5625 | 0.9996 | 0.189 ± 0.131 |
| 25           | ORF1a.Y105S | 0.9994 | 0.088 ± 0.122 |

Figure DD: We compare BVAS to MAP estimates for the selection coefficient $\beta$ for 4 values of the MAP regularization parameter $\tau$. As $\tau$ ranges across $\{2^8, 2^{10}, 2^{12}, 2^{14}\}$, the effective regularization parameter $\gamma_{\text{reg}} = \tau/\nu$ is approximately given by $\{4, 16, 64, 256\}$, since $\nu \approx 63.9$. 

31
Figure EE: We compare BVAS to MAP estimates for the growth rates of the top 10 fittest lineages (as ranked by BVAS) for 4 values of the MAP regularization parameter $\tau$. MAP growth rate estimates with $\tau \geq 10^{12}$ appear to be over-regularized towards Wuhan A. We also note that MAP uncertainty estimates are implausibly narrow.

Figure FF: In this companion figure to Figure EE, we compare BVAS to MAP estimates for the growth rates of 9 select lineages for 3 values of the MAP regularization parameter $\tau$. MAP growth rate estimates with $\tau \geq 10^{12}$ appear to be over-regularized towards Wuhan A.

Figure GG: In this companion figure to Figure EE-FF, we compare BVAS and MAP ($\tau = 2^{12}$) estimates of $R/R_A$ across 1662 PANGO lineages. If we regress MAP $R/R_A$ against BVAS $R/R_A$ we find a slope of 0.85, i.e. MAP growth rate estimates are about 15% lower than BVAS growth rate estimates.
Figure HH: We depict genome-wide Manhattan plots obtained with MAP for 4 distinct values of the regularization parameter $\tau \in \{2^8, 2^{10}, 2^{12}, 2^{14}\}$. These choices of regularization parameters lead to $\{272, 102, 68, 58\}$ alleles with effect sizes greater than 0.05, respectively, and $\{595, 275, 116, 73\}$ alleles with effect sizes greater than 0.025, respectively. MAP results obtained with $\tau \leq 2^{10}$ arguably exhibit implausibly high numbers of non-neutral alleles.

| BVAS Rank | BVAS PIP | BVAS $\beta$ | MAP Rank | MAP $\beta$ |
|-----------|----------|--------------|-----------|-------------|
| S1348K    | 1        | 1.000        | 0.371 ± 0.048 | 1       | 0.219 ± 0.020 |
| S1458K    | 2        | 1.000        | 0.430 ± 0.103 | 7       | 0.146 ± 0.027 |
| S1489K    | 3        | 1.000        | 0.386 ± 0.097 | 27      | 0.118 ± 0.025 |
| S1501Y    | 4        | 1.000        | 0.364 ± 0.119 | 2       | 0.181 ± 0.026 |
| M837T     | 5        | 1.000        | 0.356 ± 0.069 | 24      | 0.120 ± 0.026 |
| S5614G    | 6        | 1.000        | 0.339 ± 0.119 | 68      | 0.056 ± 0.009 |
| S1325Y    | 7        | 1.000        | 0.170 ± 0.058 | 61      | 0.065 ± 0.024 |
| S1701H    | 8        | 1.000        | 0.216 ± 0.103 | 6       | 0.146 ± 0.025 |
| S2477N    | 9        | 1.000        | 0.217 ± 0.146 | 28      | 0.113 ± 0.026 |
| S7851R    | 10       | 1.000        | 0.338 ± 0.146 | 40      | 0.102 ± 0.029 |
| ORF1a/K1383R | 11   | 0.986       | 0.106 ± 0.045 | 80      | 0.043 ± 0.023 |
| S1444K    | 12       | 0.969       | 0.206 ± 0.126 | 6       | 0.142 ± 0.026 |
| ORF1b/H1027Y | 13  | 0.988       | 0.185 ± 0.077 | 77      | 0.064 ± 0.025 |
| ORF1a/AS359W | 14  | 0.986       | 0.087 ± 0.039 | 71      | 0.048 ± 0.021 |
| S1655Y    | 15       | 0.964       | 0.328 ± 0.154 | 5       | 0.150 ± 0.023 |
| ORF1a/AS234V | 16  | 0.954       | 0.115 ± 0.090 | 96      | 0.031 ± 0.023 |
| S6719K    | 17       | 0.956       | 0.321 ± 0.233 | 12      | 0.131 ± 0.000 |
| ORF1a/AS235H | 18  | 0.947       | 0.186 ± 0.145 | 4       | 0.163 ± 0.028 |
| S7869N    | 19       | 0.845       | 0.030 ± 0.182 | 79      | 0.042 ± 0.028 |
| N1D1L    | 20       | 0.815       | 0.206 ± 0.198 | 87      | 0.036 ± 0.029 |
| S4420Q    | 21       | 0.883       | 0.359 ± 0.243 | 124     | 0.035 ± 0.000 |
| S1584K    | 22       | 0.832       | 0.267 ± 0.032 | 11      | 0.131 ± 0.000 |
| M832T     | 23       | 0.813       | 0.226 ± 0.063 | 13      | 0.131 ± 0.000 |
| S1571F    | 24       | 0.868       | 0.245 ± 0.299 | 31      | 0.110 ± 0.000 |
| S1998K    | 25       | 0.789       | 0.209 ± 0.315 | 14      | 0.130 ± 0.000 |

Figure II: (Left) We compare the top 25 BVAS hits to MAP ($\tau = 2^{12}$) hits, where BVAS hits are ranked by PIP. (Right) We compare the top 25 MAP ($\tau = 2^{12}$) hits to BVAS hits, where MAP hits are ranked by magnitude of effect size. We note that 12 hits are common to both top 25 rankings.
Figure JJ: We plot BVAS (top) and PyR₀ (bottom) genome-wide Manhattan plots side-by-side. Notably the effects inferred by BVAS are much sparser and the largest effects are larger in magnitude. PyR₀ results use a regularization scale $\sigma_3 = 2 \times 10^{-4}$.

Table E: In this companion table to Table 3 in the main text we report PIPs and estimated coefficients $\alpha$ for vaccination-dependent effects for an analysis in which the vaccination status is ‘vaccinated’. We report effects that are ranked in the top 75 by PIP. Estimated $\beta$ coefficients and PIPs for the corresponding vaccination-independent effects are reported in the two rightmost columns.
Figure KK: (Left) We compare the top 25 BVAS hits to PyR₀ hits, where BVAS hits are ranked by PIP. (Right) We compare the top 25 PyR₀ hits to BVAS hits, where PyR₀ hits are ranked by z-score. We note that PyR₀ uncertainty estimates are implausibly small, an unfortunate consequence of variational inference. We note that 7 hits are common to both top 25 rankings.

Figure LL: We compare BVAS and PyR₀ estimates of $R/R_A$. Uncertainty bars denote 95% credible intervals. (Left) We compare the top 10 most fit lineages (as ranked by BVAS). (Right) We compare a selection of notable lineages (as ranked by BVAS). PyR₀ uncertainty estimates tend to be much narrower, a property that results from overconfident uncertainty estimates of mutation-specific coefficients. Note that both methods infer significant diversity in B.1.1 fitness estimates (see Figure G).
Figure MM: We plot $\beta$ coefficients inferred by BVAS against coefficients inferred by PyR$_0$. Notably PyR$_0$ infers many more non-negligible coefficients. As a consequence the effect sizes of top BVAS hits tend to be considerably larger. PyR$_0$ results use a regularization scale $\sigma_3 = 2 \times 10^{-4}$.

Figure NN: We compare BVAS and PyR$_0$ estimates of $R/R_A$ across 1662 PANGO lineages. PyR$_0$ estimates are markedly higher than BVAS estimates for Omicron lineages and otherwise tend to be lower.

Figure OO: We depict allele fraction time series for N:R203K in England throughout the first two years of the pandemic. The allele fractions for ORF14:G50N and N:G204R are nearly identical. BVAS infers that all three alleles are likely passenger mutations.
Figure PP: We compare BVAS estimates from vaccination-independent analyses to estimates derived from analyses with vaccination-dependent effects. We depict allele-level Posterior Inclusion Probabilities (left) and selection coefficients $\beta$ (right); i.e. both quantities correspond to vaccination-independent effects. The vaccination status used in the top (respectively, bottom) panel is ‘fully vaccinated’ (‘vaccinated’). In both panels $N^\text{tot}_{\text{min}} = 10^4$ and $N_R = 127$ regions are included in the analysis. The strong concordance implies that the incorporation of vaccination status into the model has only moderate effects on the estimates of vaccination-independent effects and model fit as a whole.

Table F: We report inference results for the top hits in our epistasis analysis, namely those with PIPs above 0.2. These are the same hits that are depicted in the interaction network diagram in the main text. The first column reports the PIP of the pairwise interaction term in the epistasis analysis with both linear and quadratic effects. The next two columns list the corresponding Spike amino acid mutations. The next two columns report the total number of SARS-CoV-2 sequences in our dataset that carry the first (respectively, second) but not the second (respectively, first) amino acid mutation. The next column reports the total number of SARS-CoV-2 sequences in our dataset that carry both amino acid mutations. The next three columns are like the counts in the previous three columns except they report the leading PANGO lineages corresponding to the counts in the previous three columns. Note that a lineage can appear in multiple columns because there is non-negligible genetic diversity in some PANGO lineages (e.g. BA.1). The next column reports the inferred interaction effect in the epistasis analysis with both linear and quadratic effects. The next two columns report the inferred linear effects in the epistasis analysis corresponding to amino acids AA1 and AA2. The last two columns report the inferred linear effects in the default analysis corresponding to amino acids AA1 and AA2.
Figure QQ: *(Top)* We depict time series for the fraction of individuals who are fully vaccinated for 34 well-sampled regions [19]. *(Middle)* We depict allele fractions for the N501Y mutation in the S gene for the same 34 regions. *(Bottom)* We depict variant fractions for the B.1.1.7 (Alpha) variant for the same 34 regions. The rise and fall of B.1.1.7 is correlated with that of S:N501Y but not identical; for example, B.1.1.7 never became dominant in Brazil.
Figure RR: We correlate vaccination status with uptake (i.e., rising prevalence) of the S:N501Y allele as well as uptake of 3 SARS-CoV-2 variants that contain S.N501Y. In each case the vertical axis encodes the difference in allele/variant fraction from November 1st 2020 to the subsequent peak (typically in March to June 2021). The horizontal axis depicts the vaccination status at the peak or at mid-peak, i.e. when the corresponding allele/variant fraction has obtained half the value of the peak. Red dashed lines depict linear fits.
Figure SS: We depict the sensitivity of BVAS estimates of allele-level Posterior Inclusion Probabilities (left), selection coefficients $\beta$ (middle), and growth rates $R/R_A$ (right) for an analysis with linear and quadratic effects to one with linear effects only. The former analysis includes 4407 selection effects; here we subset to the $A = 2975$ linear effects that are common to both analyses. The inferred linear effects are largely the same for both analyses, except for a relatively modest number of outliers. In particular there are 5 linear effects whose PIP is substantially smaller in the epistasis analysis (S:R346K, S:S371F, S:S704L, S:V213G, and S:R408S) and one linear effect (S:T95I) whose PIP is substantially larger in the epistasis analysis (where a substantial difference means a difference larger than 0.5). As expected growth rate estimates exhibit very high concordance.

Figure TT: We depict co-occurrence posterior inclusion probabilities for the 8 interaction effects and corresponding 10 linear effects in our epistasis analysis. In other words diagonal entries correspond to standard PIPs, while off-diagonal entries correspond to the posterior probability that the corresponding pair of features co-occur in the model posterior. For example the interaction term (N501Y, P681H) co-occurs frequently with (P681H, T1027I) in the posterior.
References

[1] Lee B, Sohail MS, Finney E, Ahmed SF, Quadeer AA, McKay MR, et al. Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data. medRxiv. 2022; p. 2021–12.

[2] Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science. 2022; doi:10.1126/science.abm1208.

[3] Chipman H, George EI, McCulloch RE, Clyde M, Foster DP, Stine RA. The practical implementation of Bayesian model selection. Lecture Notes-Monograph Series. 2001; p. 65–134.

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

[5] Jankowiak M. Bayesian Variable Selection in a Million Dimensions. arXiv preprint arXiv:220801180. 2022;.

[6] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

[7] Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, et al. Pyro: Deep Universal Probabilistic Programming. J Mach Learn Res. 2019;20:28:1–28:6.

[8] Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. In: Artificial Intelligence and Statistics. PMLR; 2009. p. 73–80.

[9] Chan ER, Jones LD, Linger M, Kovach JD, Torres-Teran MM, Wertz A, et al. COVID-19 Infection and Transmission Includes Complex Sequence Diversity. bioRxiv. 2022;.

[10] Lau MS, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proceedings of the National Academy of Sciences. 2020;117(36):22430–22435.

[11] Endo A, et al. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome open research. 2020:5.

[12] Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, 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. 2020;20(8):911–919.

[13] Miller D, Martin MA, Harel N, Tirosch O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nature communications. 2020;11(1):1–10.

[14] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID’s innovative contribution to global health. Global challenges. 2017;1(1):33–46.

[15] Rambaut A, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology. 2020;5(11):1403–1407.

[16] McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, et al. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Molecular biology and evolution. 2021;38(12):5819–5824.

[17] Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics. 2021;53(6):809–816.

[18] Bissiri PG, Holmes CC, Walker SG. A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2016;78(5):1103–1130.
[19] Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Coronavirus Pandemic (COVID-19). Our World in Data. 2020;.

[20] Greaney AJ, Starr TN, Bloom JD. An Antibody-Escape Estimator for Mutations to the SARS-CoV-2 Receptor-Binding Domain. Virus Evolution. 2022;.

[21] Cubuk J, Alston JJ, Incicco JJ, Singh S, Stuchell-Brereton MD, Ward MD, et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. Nature communications. 2021;12(1):1–17.

[22] Chen Z, Pei D, Jiang L, Song Y, Wang J, Wang H, et al. Antigenicity analysis of different regions of the severe acute respiratory syndrome coronavirus nucleocapsid protein. Clinical chemistry. 2004;50(6):988–995.

[23] Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974;23(1):23–35.

[24] Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–1303.

[25] George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993;88(423):881–889.

[26] George EI, McCulloch RE. Approaches for Bayesian variable selection. Statistica sinica. 1997; p. 339–373.

[27] Dellaportas P, Forster JJ, Ntzoufras I. On Bayesian model and variable selection using MCMC. Statistics and Computing. 2002;12(1):27–36.

[28] O’Hara RB, Sillanpää MJ, et al. A review of Bayesian variable selection methods: what, how and which. Bayesian analysis. 2009;4(1):85–117.