The infinite alleles model revisited: a Gibbs sampling approach

Marc Manceau

July 22, 2021

Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
marc.manceau@ethz.ch

Abstract

The SARS-CoV-2 outbreak started in late 2019 in the Hubei province in China and the first viral sequence was made available to the scientific community on early January 2020. From there, viral genomes from all over the world have followed at an outstanding rate, reaching already more than $10^5$ on early May 2020, and more than $10^6$ by early March 2021. Phylodynamics methods have been designed in recent years to process such datasets and infer population dynamics and sampling intensities in the past. However, the unprecedented scale of the SARS-CoV-2 dataset now calls for new methodological developments, relying e.g. on simplifying assumptions of the mutation process.

In this article, I build on the infinite alleles model stemming from the field of population genetics to develop a new Bayesian statistical method allowing the joint reconstruction of the outbreak’s effective population sizes and sampling intensities through time. This relies on prior conjugacy properties that prove useful both to develop a Gibbs sampler and to gain intuition on the way different parameters of the model are linked and inferred. I finally illustrate the use of this method on SARS-CoV-2 genomes sequenced during the first wave of the outbreak in four distinct European countries, thus offering a new perspective on the evolution of the sampling intensity through time in these countries from genetic data only.

Keywords : population genetics | phylodynamics | infinite alleles | coalescent | prior conjugacy | Gibbs sampling |
Contents

1 Introduction 3

2 Model and notation 5
   2.1 Model parameters 6
   2.2 Past sampling and coalescent history 6
   2.3 Density of the full history 7

3 Inference method 7
   3.1 Observations and inference strategy 7
   3.2 Prior conjugacy properties for parameters 9
   3.3 Data augmentation with the past coalescent history 10
   3.4 Summary of the Gibbs sampler 13

4 Numerical validation of the method 13
   4.1 Past coalescent history 13
   4.2 Simulation-based calibration 14
   4.3 Running time assessment 15

5 Empirical application 16
   5.1 Data collection and preprocessing 16
   5.2 MCMC specifications 16
   5.3 Results 17

6 Discussion 19

A Details on the numerical method 25

B Case count data in the four countries 26

C Gibbs sampler code documentation 27

D Acknowledgements for data collection 28
1 Introduction

The concept of descent with modification is central in modern biology, where biological entities evolving across various spatial and temporal scales (e.g. cells, individuals, species) can be seen as atomic particles carrying molecular sequences, that are passed on to their descent, while accumulating small gradual changes. As a result, the patterns of genetic differentiation obtained in a sample of particles depend on the underlying population dynamics, and can be analysed to retrieve information on this unobserved population dynamics. This is the aim shared by two related fields called population genetics and phylodynamics.

Population genetics and phylodynamics Molecular sequences are nowadays routinely being collected and analyzed throughout the tree of life, to address a wealth of biological questions, across fields such as ecology, anthropology, macroevolution, developmental biology, or epidemiology. In this manuscript, I focus on methods designed to investigate the population dynamics of a system through the analysis of genetic polymorphism. These methods have been applied across plenty of temporal and geographical scales, e.g. in ecology to study the population size trajectory of species [Parag et al. 2021], in epidemiology to estimate the prevalence of an infectious disease from sequences of pathogens sampled during an outbreak [Stadler et al. 2013], or in paleontology to study the species diversity trajectory of a clade over macroevolutionary time-scales [Morlon et al. 2011]. While both fields address similar questions and may seem intertwined, their methodologies remained quite distinct, giving rise to two branches in the literature.

Population genetics approaches primarily aim at studying genetic variation within populations through time, based on genetic data. The recognition of the central influence of demography on genetic variation fostered the development of statistical methods aiming at inferring past demography from observed genetic polymorphism. The field has been very active since the beginning of the 70s, and most of the early theory is now digested in textbooks presenting the coalescent, with or without demography complications ([Tavare 2004; Hein et al. 2004; Durrett 2008]. This early work relied on simplifying assumptions such as the infinite alleles or infinite sites models, when genetic data has been sampled at a single point in time. Elegant analytical developments of the probability distribution of summary statistics were derived, allowing one to investigate, e.g., population growth ([Kuhner et al. 1998], population structure ([Beerli and Felsenstein 1999], selection ([McDonald and Kreitman 1991], or the presence of recombination ([Hudson 1983]. Contemporary empirical applications usually deal with more complicated demography scenarios and samples taken at multiple points in time, and have thus adopted two different strategies. First, some studies rely on Principal Component Analysis or summary statistics that have been previously derived in very simple settings ([Novembre et al. 2008]. This approach is extremely fast and appropriate for an initial exploration of the dataset, still it lacks a quantitative aspect. Second, the rise of computational power fostered the development of Approximate Bayesian Computation to fit parameter-rich models using computationally intensive procedures ([Skoglund et al. 2014; Kim et al. 2017].

Phylodynamics approaches stem from the field of phylogenetics, which aims at reconstructing the ancestral relationships between individuals, together with their evolutionary parameters, based on genetic data. In this field as well, researchers have acknowledged the key role of the demography in shaping the phylogenetic tree and hence the observed molecular patterns. This in turn promoted the rise of a subfield called phylodynamics, aiming at inferring the demography using molecular sequences, by integrating over precise phylogenetic relationships. The two main demography frameworks used in the field are (i) the coalescent, borrowed from [Kingman 1982]’s work in population genetics and (ii) birth-death processes, relying on seminal results by [Kendall 1948] and [Nee et al. 1994]. Compared to population genetics methods, there has been a cultural change towards more precise estimation relying on computationally intensive Bayesian inference methods. These methods rely on many superimposed model
layers, among which e.g. models of clock evolution \cite{Lepage2007}, models of across-locus variation \cite{Lartillot2004}, and models of molecular substitution \cite{Lanave1984}. Moreover, phylodynamic methods have been developed to take into account serially sampled molecular data \cite{Stadler2010}. Population dynamics has been modeled either in a coalescent framework using e.g. time-varying population size \cite{Pybus2000} \cite{Drummond2002} \cite{Pybus2003}, or in a birth-death framework, where the population is already free to fluctuate with constant birth and death rates, but larger variations can be allowed using time-varying parameters \cite{Morlon2011} \cite{Stadler2013}. As an alternative to time-dependent processes, some studies have attempted to introduce diversity-dependence processes in either a coalescent framework \cite{Volz2009}, or a birth-death framework \cite{Etienne2012} \cite{Leventhal2013}. Population structure can be modeled in a coalescent framework with discrete demes exchanging genes through migration \cite{Ewing2004} \cite{Vaughan2014} \cite{Muller2017}. In a birth-death process, structure is modeled using so-called multi-type birth-death processes, where different types are associated with different birth and death parameters, and individuals from a given type can either give birth to other types or directly change type \cite{Mackison2007} \cite{Beaulieu2016} \cite{Maliet2019} \cite{Barido-Sottani2020}. Finally, methods have been developed to jointly consider occurrence and molecular data. In a coalescent framework, occurrences are assumed to be the result of a Poisson sampling process among the total population \cite{Rasmussen2011} \cite{Parag2020}. In a birth-death process, an individual can be sampled and sequenced at a given rate – in which case it appears in the tree – or sampled without being sequenced at another rate – in which case it is a simple occurrence \cite{Vaughan2019} \cite{Gupta2020} \cite{Manceau2021}.

**Motivating example**  In this paper, I focus on inferring population dynamics for biological systems where (i) genetic polymorphism is sampled through time, and (ii) an ever increasing amount of sequences are being collected, challenging state-of-the-art methods for phylodynamic analysis.

The current SARS-CoV-2 pandemic provides the archetypal dataset that I propose to model. The outbreak survey started in late 2019, and the first viral genome was already published and made available for research on the 10th of January 2020. New sequences followed at an outstanding rate. By early May 2020, already more than $10^5$ viral sequences were available from across the world. A bit less than a year after, $10^6$ sequences have been reached before early March 2021. Developing statistical tools capable of keeping up with the pace of data acquisition thus represents a methodological challenge.

SARS-CoV-2 genomes have already been used to address a number of epidemiology-related questions, among which assessing the number and origins of introductions in a given locality \cite{Gonzalez-Reiche2020} \cite{Lemey2020}, the magnitude of super-spreading events \cite{Li2020}, or estimating the reproductive numbers of local outbreaks \cite{Vaughan2020}. Yet, phylogenetics/phylodynamic approaches do not scale well to large numbers of sequences and empirical applications typically require subsampling the original datasets.

The virus genome is approximately $3 \times 10^4$ nucleotides long, and its mutation rate, quite heterogeneous across the genome, has been estimated around 22 mutations per year per genome \cite{Hadfield2018}. As a result, new alleles and polymorphic sites of the genome have accumulated in the data at a slow pace. Together with the outstanding number of sequences, this rather slow mutation rate advocates for the use of simplifying assumptions of the mutation process.

**The infinite alleles model**  Phylodynamic analyses generally assume a very realistic mutation process. Sequences have a finite number of sites, and each mutation hits a randomly chosen nucleotide, with a realistic substitution process ranging from the Jukes-Cantor to the Generalized Time Reversible model. Selection might even be modeled
and nucleotides might have different mutation rates along the sequence. While these realistic models are very well designed to study fine-grain processes or processes happening over long timescales, they do not appear to be the best option to process large numbers of similar sequences. In this manuscript, I take a step back and aim at bringing back into fashion a simplifying assumption that has been traditionally considered in the early days of the neighbouring field of population genetics, namely the infinite alleles model. Under this model, each mutation hitting a sequence always creates a new allele never observed before. If we imagine that each sequence is a ball and an allele is a colour, genetic data thus simplifies as a sampling record of coloured balls through time, as illustrated in Figure 1 (Durrett 2008).

Analytical tractability is the main reason why the infinite alleles model is used nowadays. Following the past history of one genetic sequence backward in time, it can either (i) coalesce with another lineage that belongs to the same allele; or (ii) if it is the only representative of its allele, it can find the mutation that gave rise to it. Once this original mutation is found, everything else in the past is forgotten. The infinite alleles model was studied extensively during the golden age of population genetics, in combination with the coalescent model and for sequences sampled at a unique point in time. A closed-form analytical characterization of the probability distribution of the allele frequency spectrum in this setting exists, called *Ewen’s sampling formula* (Ewens 1972).

The dynamics of the colour assemblage through time is informative of the underlying population dynamics that we are interested in inferring. I propose to work under a Bayesian framework, and to rely on population dynamics and sampling process assumptions similar to what has been recently used in phylodynamics. To ensure fast convergence of the Markov Chain Monte Carlo (MCMC) method that is used for inference, the model is (i) carefully built such that data augmentation can be performed efficiently, and (ii) relies on prior conjugacy properties and Gibbs sampling moves. This approach has been successfully applied to other phylogenetics methods before, and has shown a much faster convergence of MCMC methods relying on Gibbs sampling moves as compared to Metropolis-Hastings moves (Lartillot 2006). Further, prior conjugacy properties allow one to build a better intuition on the interactions between different parameters, which proves particularly convenient for the choice of prior distributions. As compared to state of the art phylogenetics methods, I aim at integrating over the unknown ancestral relationships more efficiently, with the hope to warrant dataset analysis on a larger scale.

**Manuscript outline** In Section 2 I introduce in more details the model assumptions, before turning to the inference strategy in Section 3. I then present in Section 4 some sanity checks and validate the inference method on simulated data. An empirical application in Section 5 illustrates the use of the method on the SARS-CoV-2 sequences sampled during the first wave hitting Europe in 2020. Finally, I discuss in Section 6 the results of this paper as well as the future research challenges it opens. This manuscript is released along with the code implementing the method, and details on the implementation and use of the code are provided in Supp. Mat. C.

## 2 Model and notation

I build here on work by Parag et al. (2021) and Karcher et al. (2020), who both consider a sampling process on top of a coalescent model with piecewise-constant effective population sizes. The coalescent process is very conveniently described backward in time, and time will thus be, throughout the manuscript, the calendar time before present, in units of days for empirical applications, with $t = 0$ at present and $t \to \infty$ in the past.
2.1 Model parameters

The model is built around the following four key parameters.

First, the past effective population size is piecewise-constant on a partition of \((0, +\infty)\) into \(p\) successive disjoint intervals \((\Delta_j^{(N)})_{j=0}^{p-1}\) delimited by 0, \(\infty\), and the \(p-1\) times \((t_j^{(N)})_{j=1}^{p-1}\), i.e.

\[
N_t = \sum_{j=0}^{p-1} N_j \mathbb{1}_{\Delta_j^{(N)}(t)} ,
\]

where \(\Delta_j^{(N)} = [t_j^{(N)}, t_{j+1}^{(N)})\), with convention \(t_0^{(N)} = 0\) and \(t_p^{(N)} = \infty\).

Second, following Parag et al. (2021), the past sampling intensity is also a piecewise-constant function, on a possibly different partition \((\Delta_j^{(S)})_{j=0}^{p'-1}\) of \((0, \infty)\), delimited by 0, \(\infty\), and the \(p' - 1\) times \((t_j^{(S)})_{j=1}^{p'-1}\), i.e.

\[
S_t = \sum_{j=0}^{p'-1} S_j \mathbb{1}_{\Delta_j^{(S)}(t)} ,
\]

where \(\Delta_j^{(S)} = [t_j^{(S)}, t_{j+1}^{(S)})\), with convention \(t_0^{(S)} = 0\) and \(t_{p'}^{(S)} = \infty\).

Last, the mutation rate \(\mu\) and generation time \(g\) are constant through time.

In a Bayesian framework, these parameters are random variables which are assigned prior distributions. The effective population sizes \((N_j)_{j=0}^{p-1}\) are assumed to be \textit{a priori} distributed according to a Generalized Inverse Gaussian distribution, while the sampling intensities \((S_j)_{j=0}^{p'-1}\) are assumed to be \textit{a priori} distributed according to a Gamma distribution. Finally, the mutation rate \(\mu\) and generation time \(g\) are respectively \textit{a priori} distributed according to a Gamma and Inverse-Gamma distribution. The choice of these prior distributions will be explained in Section 3 when discussing the posterior inference of these variables.

2.2 Past sampling and coalescent history

We assume that the sampling history is given by a Poisson Point Process (PPP) with rate

\[
\lambda_t^{(b)} := S_t N_t
\]

generating the set of ordered sampling times \(B = (b_i)_{i=0}^{B-1}\) of all individuals. The total number of sampling events is denoted \(B\), and individuals are numbered from 0 to \(B - 1\) in reverse birth time order. We will also call these sampling times the \textit{birth times} of lineages when considering the history backward in time (hence the name \(B\)). Lineages begin their backward-in-time journey as singletons \(\{i\}\) where \(i\) corresponds to the individual’s number.

The past history of these lineages is further assumed to follow a standard coalescent with effective population size \(N_t\), generation time \(g\) and differentiation under an infinite alleles model with mutation rate \(\mu\). That is, while there are \(k_t\) lineages alive in the process, the next coalescent (resp. differentiation) event happens with rate,

\[
\lambda_t^{(c)} := \left(\frac{k_t}{2}\right) (g N_t)^{-1} \quad \text{and} \quad \lambda_t^{(d)} := \mu k_t
\]
When there is a coalescent event, two lineages $L_i$ and $L_j$, uniformly sampled among the $k_t$ living lineages at that time, are merged together in a unique lineage $L_i \cup L_j$. When there is a differentiation event, one of the $k_t$ living lineages is uniformly chosen to be killed. Forward in time, a coalescence corresponds to an individual giving birth to another individual, whereas a differentiation event corresponds to the acquisition of an original mutation responsible for the creation of an allele.

The past coalescent history thus generates in particular a partition of individuals into an allele partition $A$ corresponding to the collection of all lineages killed by a mutation. It also generates the times at which differentiation and coalescent events — jointly referred to as death events — happened in history. In order to record these, we take the following approach. Lineages are initially numbered with the same number as the individual’s number it carries. Each coalescence involving two lineages numbered $j < i$ at time $t$ is considered to kill lineage $i$, and to keep living in lineage $j$ (see arrows on Fig. 1). By a slight abuse of language, we call such an event the death of individual $i$, and the time at which a mutation is found is the death time of the very first individual of the allele (see crosses on Fig. 1).

The coalescent history of all individuals is recorded in $H = (h_i, o_i)_{i=0}^{B-1}$, where $h_i$ is the death time of individual $i$ and $o_i \in \{0, 1, \ldots, i\}$ is the output of the death event, i.e. the number $j < i$ of the lineage in which lineage $i$ is merged if there is a coalescence, or $o_i = i$ if there is a mutation. The total number of alleles is denoted $D = \sum_{i=0}^{B-1} \mathbb{1}_{o_i = i}$. Finally, the record of the $B$ birth times $(b_i)$ and death times $(h_i)$, together with boundaries $t = 0$ and $t = \infty$, yield a partition of the timeline into $2B + 1$ successive intervals that are denoted $(\Delta_i)_{i=0}^{2B}$. On any such interval $\Delta_i$, the number of lineages remain constant and is denoted $k_i$.

### 2.3 Density of the full history

Figure 1 summarizes all notation introduced so far. Note already that, knowing $B$ and $H$, i.e. the full sampling and coalescent history, is enough to know the partition $A$ of individuals into alleles. The density of this full past history $B, H$ given $N, S, \mu, g$ is further given by

$$P(B, H | N, S, \mu, g) = \left( \prod_{i=0}^{B-1} \lambda_{b_i}^{(b)}(\lambda_{h_i}^{(c)} \mathbb{1}_{o_i \neq i} + \lambda_{h_i}^{(d)} \mathbb{1}_{o_i = i}) \right) \exp \left( - \int_0^\infty (\lambda_{t}^{(b)} + \lambda_{t}^{(c)} + \lambda_{t}^{(d)}) dt \right)$$

(2.4)

The above density belongs to the exponential family, and thus lends itself well to inference via a Gibbs sampling strategy, with priors that belong to other exponential family distributions. In the next Section, I turn to the description of this inference method.

### 3 Inference method

#### 3.1 Observations and inference strategy

Data consist in the observation of the sampling times $B$, together with the partition $A$ of the set of $B$ individuals into $D$ alleles. I aim at inferring the posterior distribution of $N, S, \mu, g$, which consists, in a Bayesian framework, in sampling from,

$$P(N, S, \mu, g | A, B) = \int_H P(N, S, \mu, g, H | A, B)$$


Figure 1: Model notation. Time is oriented from the present towards the past. To the left, individuals, represented as circles at their sampling time, are numbered in decreasing sampling times order and are colored according to the allele they belong to. Past coalescent events are represented with arrows, differentiation events are represented with crosses. The past coalescent history leads to the allele partition \( A = \{\{0, 1, 4\}, \{2, 3\}\} \). To the right, the parameters of the model that we are interested in inferring: the piecewise-constant past effective population sizes \( N \), the piecewise-constant past sampling intensities \( S \), the constant mutation rate \( \mu \), and the constant generation time \( g \).

where the right-hand side is augmented using the past coalescent history.

I take a Gibbs sampling approach to design a MCMC which converges to the stationary distribution of the augmented target distribution \( \mathbb{P}(N, S, \mu, g, H \mid A, B) \). To do so, I derive efficient ways to alternatively sample from the following conditional laws,

\[
\begin{align*}
\mathbb{P}(N_j \mid N_{-j}, S, \mu, g, B, H) \\
\mathbb{P}(S_j \mid N, S_{-j}, \mu, g, B, H) \\
\mathbb{P}(\mu \mid N, S, g, B, H) \\
\mathbb{P}(g \mid N, S, \mu, B, H) \\
\mathbb{P}(H_i, O_i \mid N, S, \mu, g, A, B, H_{-i}, O_{-i})
\end{align*}
\]

where \( N_{-j}, S_{-j} \) denote the effective population sizes and sampling intensities over all intervals other than the one numbered \( j \), and \( (H_{-i}, O_{-i}) \) denotes the death information of all individuals other than the one numbered \( i \). Remark here that, on the first three lines, \( A \) disappeared from the conditioning because \( A, B, H = B, H \), i.e. knowing the sampling times and the coalescent history is enough to know the allele partition.

Designing an efficient Gibbs sampler in this context relies on the optimization of two critical steps: (i) one needs to derive the posterior distribution of the parameters conditioned on the augmented data; and (ii) one needs to efficiently perform a data augmentation step, i.e. simulate the past history \( H \) conditioned on the observed data and parameter values.
3.2 Prior conjugacy properties for parameters

3.2.1 Effective population size

Recall that $N_j$ is a priori independent on $N_{-j}, S, \mu, g$ and is distributed according to a Generalized Inverse Gaussian distribution denoted $\mathcal{IG}(\lambda, \chi, \psi)$. The following will justify this choice of prior. The GIG distribution belongs to the exponential family and is characterized by its density, usually parameterized as,

$$P(N_j) \propto N_j^{\lambda-1} \exp \left( -\frac{1}{2}(\chi N_j^{-1} + \psi N_j) \right).$$

Its posterior is thus given by,

$$P(N_j | N_{-j}, S, \mu, g, B, \mathcal{H}) \propto P(N_j) P(B, \mathcal{H} | N, S, \mu, g)$$

$$\propto N_j^{\beta-1} \exp \left( -\frac{1}{2}(\chi N_j^{-1} + \psi N_j) \right)$$

$$\sum_{i=0}^{B-1} 1_{b_i \in \Delta_j^{(N)}} - 1_{b_i \notin \Delta_j^{(N)}}$$

$$\exp \left( -N_j^{-1} g^{-1} \sum_{l=0}^{2B} \left( k_l / 2 \right) |\Delta_l \cap \Delta_j^{(N)}| - N_j \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}| \right),$$

where the last line is obtained again by substituting the density of $B, \mathcal{H}$ using Equation (2.4), before dropping out all terms which do not depend on $N_j$.

This proves that the prior and posterior are conjugate distributions, with

$$N_j | N_{-j}, S, \mu, g, B, \mathcal{H}$$

$$\sim \mathcal{IG} \left( \lambda + B(\Delta_j^{(N)}) - C(\Delta_j^{(N)}), \chi + g^{-1} \sum_{l=0}^{2B} k_l (k_l - 1) |\Delta_l \cap \Delta_j^{(N)}|, \psi + 2 \sum_{k=0}^{p'-1} S_k |\Delta_k^{(S)} \cap \Delta_j^{(N)}| \right), \hspace{1cm} (3.5)$$

where $C(\Delta_j^{(N)})$ and $B(\Delta_j^{(N)})$ are respectively the number of coalescent and sampling events happening over interval $\Delta_j^{(N)}$.

The choice of a conjugate prior will help (i) simplify the Gibbs sampling process, and (ii) provide a better intuitive understanding of the factors that influence the distribution of $N$.

3.2.2 Sampling intensity

Recall that $S_j$ is a priori independent on $N, S_{-j}, \mu, g$ and that $S_j \sim \Gamma(\alpha, \beta)$. Its posterior is thus given by,

$$P(S_j | N, S_{-j}, \mu, g, B, \mathcal{H}) \propto P(S_j) P(B, \mathcal{H} | N, S, \mu, g)$$

$$\propto S_j^{\alpha-1} \exp (-\beta S_j) \sum_{i=0}^{B-1} 1_{b_i \in \Delta_j^{(S)}} \exp \left( -S_j \sum_{k=0}^{p'-1} N_k |\Delta_k^{(N)} \cap \Delta_j^{(S)}| \right)$$

where the last line is obtained again by substituting the density of $B, \mathcal{H}$ using Equation (2.4), before dropping out all terms which do not depend on $S_j$. 

9
This shows that the prior and posterior are conjugate distributions, with
\[
S_j | N, S_{-j}, \mu, g, B, \mathcal{H} \sim \Gamma \left( \alpha + B(\Delta_j^{(S)}), \beta + \sum_{k=0}^{p-1} N_k | \Delta_k^{(N)} \cap \Delta_j^{(S)} \right).
\] (3.6)

3.2.3 Mutation rate

Recall that, \textit{a priori}, \( \mu \sim \Gamma(\alpha, \beta) \). Following the same strategy as above, its posterior is given by,
\[
P(\mu | N, S, g, B, \mathcal{H}) \propto P(\mu, B, \mathcal{H} | N, S, g) \\
\propto P(\mu) P(B, \mathcal{H} | N, S, \mu, g) \\
\propto \mu^{a-1} \exp(-\beta \mu) \mu^D \exp \left( -\mu \sum_{l=0}^{2B-1} k_l | \Delta_l | \right).
\]

We conclude that the prior and posterior of \( \mu \) are conjugate distributions, with,
\[
\mu | N, S, g, B, \mathcal{H} \sim \Gamma \left( \alpha + D, \beta + \sum_{l=0}^{2B-1} k_l | \Delta_l | \right).
\] (3.7)

3.2.4 Generation time

Last, recall that \( g \) is \textit{a priori} distributed according to an Inverse-Gamma distribution denoted \( \Gamma^{-1}(\alpha, \beta) \). Its posterior is given by,
\[
P(g | N, S, \mu, B, \mathcal{H}) \propto P(g) P(B, \mathcal{H} | N, S, \mu) \\
\propto g^{-(\alpha+1)} \exp(-\beta g^{-1}) \prod_{i \neq i}^{n-1} g^{-\frac{2B}{2}} \sum_{i=0}^{p-1} \left( \frac{k_i}{2} \right) N^{-1} | \Delta_i \cap \Delta_j^{(N)} |
\]
which means that the posterior is
\[
g | N, S, \mu, B, \mathcal{H} \sim \Gamma^{-1} \left( \alpha + B - D, \beta + \sum_{l=0}^{2B-1} \sum_{j=0}^{p-1} \left( \frac{k_j}{2} \right) N_j^{-1} | \Delta_l \cap \Delta_j^{(N)} | \right).
\] (3.8)

This ends the description of the four conjugate priors used for \( N, S, \mu, g \). They will be used in the final Gibbs sampler to quickly update the posterior of these variables of interest, provided \( B, \mathcal{H} \) are observed.

3.3 Data augmentation with the past coalescent history

Under the assumption of an infinite alleles model with constant effective population size \( N \) and constant mutation rate \( \mu \), the distribution of the past history of a sample of \( n \) genes taken at a single point in time can be efficiently sampled using the equivalence with the Hoppes urn process (Durlett 2008). Yet, when sequences are heterochronous
i.e. have been sampled through time instead of at a single point in time – and when the population size is not constant anymore, an alternative strategy is needed.

Recall that \( H_i \in \mathbb{R}^+ \) is the death time of a focal individual \( i \), and that \( O_i \in \{0, 1, \ldots, i\} \) is the death output. Death can happen either through coalescence with other lineages \((O_i = j < i)\), or through mutation \((O_i = i)\). We aim here at sampling death information \((H_i, O_i)\) of each sequence in turn, conditioned on everything else, i.e. when \( N, S, \mu, g, A, B, H_{-i}, O_{-i} \) are fixed.

Figure 2: Gibbs sampling approach for simulating the past history of a sample given the allele partition. A-F. Each leaf is considered in turn, and its time of extinction is sampled given the rest of the coalescent history is fixed.

In a standard coalescent process, one can draw \((H_i, O_i)\) from left to right: if a focal lineage \( i \) sees \( k_i \) other lineages already drawn to the left, at time \( t \), it coalesces with one of them at rate \( k_i (N t g)^{-1} \), or finds a mutation at rate \( \mu \). Yet, when imposing our conditioning, only a fraction of these events leads to the known allele partition \( A \) and is compatible with the known coalescent history \((H_{-i}, O_{-i})\). In particular, the allele partition and past coalescent history excluding \( i \) reveal that (i) lineage \( i \) dies through a coalescent event if and only if it can coalesce with an other lineage to the left, and (ii) lineage \( i \) cannot die before being sampled, nor before reaching the very last coalescence of other lineages with itself, i.e. \( H_i \geq m_i := \max (\{b_i\} \cup \{H_j, j > i \text{ such that } O_j = i\}) \).

A first strategy to draw \((H_i, O_i)\) conditioned on everything else would thus consist in drawing \((H_i, O_i)\) without conditioning, while subsequently rejecting simulations which outcome is incompatible with \( A, O_{-i}, H_{-i} \). I take here
the following alternative approach to avoid rejecting too many simulations. On all intervals $\Delta_i = (t_i, t_{i+1})$ where the total number of lineages $k_i$ remains constant and where $t_i \geq m$, the probability that $H_i$ falls within the interval conditioned on everything else is computed,

$$
P(H_i \in \Delta_i \mid A, B, H_{-i}, O_{-i}) \propto P(H_i \in \Delta_i, A \mid B, H_{-i}, O_{-i})
$$

$$
= \left[ P(H_i \in \Delta_i \mid B, H_{-i}, O_{-i}) P(A \mid B, H_{-i}, O_{-i}) \right] \sum_{O_i} \left[ P(O_i \mid A, B, H_{-i}, O_{-i}) \right]
$$

$$
= \frac{P(H_i > t_i, H_i < t_{i+1} \mid B, H_{-i}, O_{-i}) \sum_{O_i} \left[ P(O_i \mid A, B, H_{-i}, O_{-i}) \right]}{P(H_i > t_i \mid B, H_{-i}, O_{-i}) P(H_i < t_{i+1} \mid B, H_{-i}, O_{-i})}
$$

$$
= \frac{\left(1 - e^{-\left(\mu + k_i(Ng)^{-1}\right)(t_{i+1} - t_i)}\right) N g\mu \mathbb{1}_{a_i = i} + k_i^{<i} \mathbb{1}_{a_i \neq i}}{N g\mu + k_i}
$$

(3.9)

where $k_i^{<i}$ denotes the number of living lineages to the left of $i$ belonging to the same allele $a$ as $i$.

The above formula is used to recursively compute from bottom to top the weights $w_i$ associated to all intervals $\Delta_i$ above $m_i$. Once these have been computed, we have access to, and can sample from, $P(H_i \in \Delta_i \mid A, B, H_{-i}, O_{-i}) = w_i / \sum_{\Delta_i} w_i$. Note in addition that, if lineage $i$ satisfies $o_i \neq i$, then only intervals from time $m_i$ up to time $M_i := \max H_{-i}$ matter, since it must coalesce before the mutation is found. If lineage $i$ dies through a mutation, i.e. $o_i = i$, then intervals from time $m_i = \max H_{-i}$ up to $\infty$ must be weighted, since the lineage cannot die before everybody else has coalesced into itself.

Once the interval on which the death event happens has been drawn, it remains to draw from $P(H_i \in dt \mid H_i \in \Delta_i)$, which corresponds to an exponential distribution with rate $(\mu + k_i(Ng)^{-1})$ conditioned on happening on $\Delta_i$. Further, $O_i = i$ if the lineage must die through a mutation, or $O_i$ is a uniformly chosen lineage $j < i$, in the same allele as lineage $i$, living on interval $\Delta_i$. The procedure for drawing $(H_i, O_i)$ is summarized in Algorithm 1 and illustrated in Figure 2.

Algorithm 1 Drawing a realization of $P(H_i, O_i \mid A, B, H_{-i}, O_{-i}, N, S, \mu, g)$

Input:
- Observed allele partition and sampling history $(A, B)$,
- Focal lineage number $i$ and fixed coalescent history for all other lineages $(H_{-i}, O_{-i})$,
- Parameters $N, S$ through time and $\mu, g$ fixed.

Output: A realization of $(H_i, O_i)$ conditioned on everything else.

1. Compute

$$
m_i = \max \{ \{b_i\} \cup \{H_j, j > i \text{ such that } O_j = i\} \}
$$

and

$$
M_i = (\max H_{-i}) \mathbb{1}_{a_i \neq i} + \infty \mathbb{1}_{a_i = i}
$$

2. Explore intervals, from time $m_i$ up until $M_i$ and for each interval $\Delta_i$:

   (a) record the count of $k_i$, the total number of lineages alive on the interval,
   (b) record the list $L_a^{<i}$ of all lineages $j < i$ from the same allele $a$ as lineage $i$ alive in the interval,
   (c) record the weight $w_i$ using Equation (3.9).

3. Draw the interval in which the event occurs, and then the specifics of the event:

   (a) Draw the interval $\Delta_i$ according to the weights computed at the previous step,
   (b) Draw $H_i$, from an exponential distribution with rate $\mu + k_i(Ng)^{-1}$ conditioned on happening on $\Delta_i$,
   (c) Draw $O_i$, a uniformly chosen lineage in the list $L_a^{<i}$.
3.4 Summary of the Gibbs sampler

The final Gibbs sampler is a simple MCMC that is initialized using the priors of \(N, S, \mu, g\), before relying on repeated updates of the four variables \(N, S, \mu, g\), \(\mathcal{H}\), using their conditional probabilities. Algorithm 2 summarizes these steps.

Algorithm 2 Gibbs sampling of the target distribution \(\mathbb{P}(N, S, \mu, g | A, B)\)

**Input:**
- Observed allele partition and sampling history \((A, B)\),
- Number of steps \(n\) of the MCMC,
- Prior hyperparameters \((\lambda, \chi, \psi)\) for \(N\), \((\alpha_S, \beta_S)\) for \(S\), \((\alpha_\mu, \beta_\mu)\) for \(\mu\), \((\alpha_g, \beta_g)\) for \(g\).

**Output:** A posterior sample \((N^{(i)}, S^{(i)}, \mu^{(i)}, g^{(i)})_{i=1}^n\).

1. Initialization:
   1. Fix \(\forall i \in \{0, 1, \ldots, B - 1\}\), \(H_i^{(0)} = b_i \) and \(O_i^{(0)} = \min\{j \in a_i\}\).
   2. Draw \(\forall j \in \{0, 1, \ldots, p - 1\}\), \(N_j^{(0)} \sim GIG(\lambda, \chi, \psi)\).
   3. Draw \(\forall j \in \{0, 1, \ldots, p' - 1\}\), \(S_j^{(0)} \sim \Gamma(\alpha_S, \beta_S)\).
   4. Draw \(\mu^{(0)} \sim \Gamma(\alpha_\mu, \beta_\mu)\).
   5. Draw \(g^{(0)} \sim \Gamma^{-1}(\alpha_g, \beta_g)\).
2. Chain: for \(l\) going from 1 to \(n\),
   1. Use Algorithm 1 to update, \(\forall i \in \{0, 1, \ldots, B - 1\}\), \(H_i^{(l)}(O_i^{(l)})\) using \(N^{(l-1)}, S^{(l-1)}, \mu^{(l-1)}, g^{(l-1)}\).
   2. Use Equation 3.5 to update \(N\), i.e. draw \(\forall j \in \{0, 1, \ldots, p - 1\}\), \(N_j^{(l)}\) using \(H^{(l)}, S^{(l-1)}, g^{(l-1)}\).
   3. Use Equation 3.6 to update \(S\), i.e. draw \(\forall j \in \{0, 1, \ldots, p' - 1\}\), \(S_j^{(l)}\) using \(H^{(l)}, N^{(l)}\).
   4. Use Equation 3.7 to update \(\mu\), i.e. draw \(\mu^{(l)}\) using \(H^{(l)}\).
   5. Use Equation 3.8 to update \(g\), i.e. draw \(g^{(l)}\) using \(H^{(l)}, N^{(l)}\).
   6. Record \((N^{(l)}, S^{(l)}, \mu^{(l)}, g^{(l)})\).

In the next Section, I validate the behaviour of this MCMC on simulated datasets before applying it to empirical datasets in Section 5.

4 Numerical validation of the method

4.1 Past coalescent history

I first aim to validate the procedure described in Section 3.3 and the implementation of Algorithm 1. To do so, I wrapped Algorithm 1 in a minimalist Gibbs sampler aiming at sampling from \(\mathbb{P}(\mathcal{H} | A, B, N, S, \mu, g)\). I fix \(B, A\) as well as all parameters of the model \(N, S, \mu, g\). I then use a simplified version of the MCMC described as Algorithm 2 where all updates concerning \(N, S, \mu, g\) are skipped. To draw one observation of \(\mathcal{H}\), I performed \(n = 50\) complete cycles of leaf updates, and kept only the last state reached by the MCMC.

I compare the distribution obtained using this procedure with the distribution of \(\mathcal{H}\) obtained by a naive rejection algorithm, consisting in simulating the coalescent process backwards in time, while rejecting outcomes that do not satisfy \(A\). The distributions are compared based on summary statistics computed on \(10^4\) samples drawn in both distributions: (i) the proportion of samples having a mutation in successive death events, as well as (ii) the distribution of death times. Note that the dataset must be small, for the rejection algorithm very quickly becomes computationally too intensive to be used. Figure 3 illustrates the perfect agreement between both distributions,
on a toy dataset with $N = 10$, $g = 0.1$, $\mu = 1.5$, and two alleles respectively joining individuals sampled at times $(0, 0.2, 0.5)$ and $(0.3, 0.7)$.

![Figure 3: Comparison of the distribution of $H$ conditioned on $A$ obtained using a naive rejection algorithm or a custom Gibbs sampler relying on Algorithm 1. The number of the death event in x-axis refers to an order by increasing death time. A. Proportions of mutations in each death event. B. Distributions of death times.](image)

### 4.2 Simulation-based calibration

The MCMC implementation of Algorithm 2 is further validated against simulated data using the Simulation-Based Calibration (SBC) method described by Talts et al. (2018) and summarized hereafter.

First, priors are fixed as follows, with hyperparameter values ensuring that the number of samples remains relatively small. The effective population size is piecewise-constant over $p = 2$ intervals, with $N_0, N_1 \sim GIG(\lambda = 4, \chi = 0, \psi = 0.08)$ respectively over intervals $(0, 20)$ and $(20, +\infty)$. The sampling intensity is piecewise-constant over $p' = 3$ intervals, with $S_0, S_1$ having a $\Gamma(4, 1000)$ prior respectively over intervals $(0, 10)$ and $(10, 40)$ and $S_3 = 0$ on $(40, +\infty)$. The mutation rate and generation time are respectively assigned a prior $\mu \sim \Gamma(4, 400)$ and $g \sim \Gamma^{-1}(10, 10)$.

Second, $10^4$ parameter sets are sampled from these distributions and for each parameter set, the sampling history $B$ as well as the allele partition $A$ are sampled according to the model.

Third, for each simulated dataset, the posterior distribution of $N, S, \mu, g$ conditioned on $A, B$ is sampled using the Gibbs sampler described as Algorithm 2 while using the same priors that were used for the simulation. I ran the MCMC for a total of $10^4$ steps, discarded a burn-in of $10^3$ steps at the beginning and recorded one state every 100 steps over the remaining steps.

Finally, Figure 4A shows the proportion of datasets $p_\alpha$ for which a credible interval with level $\alpha$ of the posterior distribution contains the true simulated data, for 9 values of $\alpha$ evenly spaced on $(0, 1)$. The good match between $p_\alpha$ and $\alpha$ indicates that the MCMC correctly samples the posterior distribution. This is further confirmed in Figure 4B, showing the histogram of the rank statistic associated with the $10^4$ experiments. Here, the rank statistic associated to one experiment refers to the number of samples from the posterior being less than the true value. Under the null hypothesis that we are sampling the true posterior distribution, the histogram of rank statistics should be uniform, as illustrated in Figure 4B (Talts et al. 2018).
Figure 4: Results of the SBC analysis described in the main text, validating the Gibbs sampler. A. Proportion of datasets which $100\alpha\%$ posterior credible interval recovers the true parameter, for different values of $\alpha$. B. Histogram of rank statistics. Horizontal dashed lines indicate the 99% CI of the bar heights if the histogram is uniform.

4.3 Running time assessment

I now turn to an assessment of the running time of the method on realistic datasets. I simulated datasets using a timeline cut into $p = p' = 5$ intervals with varying $N$ and $S$ values, while fixing the same hyperparameters on all intervals, so as to get datasets with total sample size $B$ regularly spaced on a log scale between 10 and 2500. On each of these, I ran $10^4$ steps of the Gibbs sampler, discarded the first $10^3$ steps, and recorded the running time and posterior samples.

Let us focus first on the running time of each of the different updates. Since for each of $B$ sequences, the update of the death event requires to compute the weights associated to a number of intervals of the order of $B$, the update of the coalescent history is expected to scale in $O(B^2)$ and to be the bottleneck of the MCMC. Figure 5A presents estimates of the running time depending on $B$, using the code released along this article on a laptop. It confirms that the update of the coalescent history – and so, a step in the MCMC – scales in $O(B^2)$.

Running a MCMC moreover requires to perform these updates repeatedly during a certain number of steps, in order to (i) escape the burn-in phase of the MCMC; and (ii) collect enough samples from the posterior to characterize it. When the samples collected through time are highly correlated, it is said that the chain is *mixing poorly*, and more steps are typically required. In order to provide a rough idea of the expected mixing behaviour on simulated datasets, I computed estimates of the effective sample size of $(N_j)_{j=0}^{p-1}$, $(S_j)_{j=0}^{p'-1}$, $\mu$, $g$. The results are shown in Figure 5B. Combined with the running time assessment, it conveys the idea that the current implementation, without further approximation or numerical optimization will likely not be useful to process more than $\sim 10^4$ sequences.

Having in mind the rough behaviour of the method, I illustrate its use in the next Section on empirical datasets encompassing $\sim 10^3$ sequences.
Figure 5: Estimating the running time of the method on simulated datasets. A. The running time of one update of each of the four variables shows that the update of the coalescent history is the bottleneck, that scales in $O(B^2)$. B. The ESS of $\mu, g$ and each of the piecewise-constant values of $S$ and $N$ as a function of the dataset size suggests that reasonable (i.e. $> 100$) ESS values for datasets up to $B \sim 10^3$ can be obtained using $\sim 10^4$ MCMC steps.

5 Empirical application

5.1 Data collection and preprocessing

Sequences collected in Switzerland (CH), Germany (DE), France (FR) and Italy (IT) between December 1st 2019 and June 1st 2020 have been downloaded on the GISAID website on the 7th of June 2021, while requiring the three options of (i) complete sequences, (ii) high coverage sequences, and (iii) complete collection date. I additionally downloaded GISAID reference genome Wuhan/WIV04/2019 (accession id EPI_ISL_402124) for use in the data preprocessing steps. All people involved in sequence data collection are acknowledged in Supp. Mat. D.

Additional removal of sequences was performed on the basis of either one of these two criteria being fulfilled, (i) sequences were not extending from position 250 to 29700 of the reference genome, (ii) the proportion of ambiguously resolved bases was higher than 10%. The remaining 1284 sequences from CH (1673 in DE, 1919 in FR, 1314 in IT) were further classified into 627 alleles (886 in DE, 1166 in FR, 750 in IT) using a custom pipeline available as part of the provided code and detailed in Supp. Mat. A.

5.2 MCMC specifications

The timeline of the analysis is fixed for $N$ and $S$. It extends from the 1st of June 2020 to the 13th of January 2020, cut into successive intervals of 4 weeks each. I assume that $\mu$ and $g$ are known from other studies and are not the focus of the inference. The mutation rate $\mu$ is fixed to 0.065 mutations per genome per day, corresponding to $8 \times 10^{-4}$ mutations per nucleotide per year, and the generation time $g$ is fixed to 5 days.

Finally, hyperparameters values for $N$ and $S$ are chosen so as to not be too informative, using a quick back-of-the-envelope reasoning around Equations (3.5) and (3.6). We imagine what could happen over a time period with few data, as happens in the beginning of the dataset. If the order of magnitude of $N$ is approximately $10^4$, and we roughly believe that one out of $5 \times 10^3$ individuals is sequenced in reality, because each individual
lives for 5 days in the model, it corresponds to sequencing one out of $2.5 \times 10^4$ individuals, i.e. $S \sim 4 \times 10^{-5}$. If we now imagine observing this situation over a time period $\Delta \sim 10$ days, we would observe on expectation 4 births. Using Equation (3.5) helps fix the prior for $(N_j)_{j=0}^{p-1}$ values, which are $GIG$ distributed with hyperparameters $(\lambda, \chi, \psi) = (4, 0, 8 \times 10^{-4})$. Then, using Equation (3.6) helps fix the prior for $(S_j)_{j=0}^{p'-1}$ values, which are Gamma distributed with hyperparameters $(\alpha_S, \beta_S) = (4, 10^5)$.

I ran the Gibbs sampler for $10^4$ steps, discarded the first $10^3$ steps as burn-in and used the remaining $9 \times 10^3$ steps for posterior inference. The code provided along with this article generates the traces and auto-correlation functions of all parameters. These are inspected visually and the ESS values of all parameters are higher than 100.

### 5.3 Results

The output of the Gibbs sampler on each country is a posterior sample of $(N_j)_{j=0}^{p-1}$ and $(S_j)_{j=0}^{p'-1}$ values through time over the fixed timeline. Figure 6 illustrates the input data (to the left) together with the output posteriors of $N$ and $S$ (to the right). In particular, the trajectories of $N$ and $S$ tell a similar story in the four countries. The outbreak slowly started in early 2020 and reached its peak around March-April 2020, before quickly decreasing in May 2020. France appear on this Figure to have experienced the longest outbreak. In all four countries, the sampling intensity increased prior to the peak of the epidemic and slowly decreased thereafter.
Figure 6: Results of the inference on SARS-CoV-2 data in four European countries during the first wave of the epidemic. CH: Switzerland, DE: Germany, FR: France, IT: Italy. First panel to the left, the raw data consists in an allele partition of sequences sampled through time, where different colors correspond to different alleles. Second and third panels, the posterior sample of $N, S$, with the median shown in green and a 95% posterior interval shown in gray.
6 Discussion

In this manuscript, I have presented a new Bayesian inference method relying on a Gibbs sampler to sample the posterior distributions of effective population sizes and sampling intensities through time, from the knowledge of an allele partition of sampled-through-time sequences. It relies on a coalescent framework under the infinite alleles model from population genetics, with constant mutation rate, constant generation time, and piecewise-constant effective population sizes and sampling intensities through time. Applying the method to SARS-CoV-2 sequences sampled throughout the first wave of the outbreak in Europe allows us to obtain estimates of effective population sizes, sampling intensities and mutation rate from genetic data only.

Inferred dynamics on empirical data  In studies relying on a coalescent model, the usage is to focus on the inferred patterns of change through time rather than on the parameter values per se. The rationale explaining this is that values might be affected in complicated ways by various model misspecifications, among which, for example, having particular distributions of offspring numbers at reproduction, having population structure not taken into account, having a different generation time or mutation rate, or having different priors on \( N \) and \( S \). Yet, as long as these model misspecifications do not change through time, one can trust the dynamics of the effective population size \( N_t \) to reflect the dynamics of the true population size. Although I have followed this convention so far throughout the manuscript, one can notice here that the orders of magnitude of inferred effective values seem nevertheless to roughly match what is known from the case count record only (shown in Supp. Mat. B).

In terms of dynamics as well, the patterns of effective population sizes appear compatible with what is known about the first wave of the outbreak in these countries. The number of infected individuals starts increasing exponentially early in 2020, reached a peak in March-April, before decreasing in response to policy measures taken – different versions of a lockdown have been implemented in the four countries around this time. Interestingly, the sampling intensity, on the other hand, starts increasing earlier during the epidemic, when the research effort started growing around SARS-CoV-2 genetics but the case counts where not so high yet. It then decreased again following the peak of the outbreak.

Opening future modeling opportunities with allele data  Overall, the results on empirical data illustrates the use and applicability of the method on large real-world datasets. Yet, the current model formulation still lacks a few realistic ingredients before it can be used to learn new aspects of an epidemic.

First and foremost, it does not take into account population structure, a feature that is likely to be present in the empirical dataset and violates the model assumptions. Further work is needed to properly incorporate different demes characterized by different population sizes and sampling intensities, exchanging sequences through migrations among demes. This will likely be the subject of future work aiming to infer population structure and population dynamics using the allele partition. Integrating over the hidden coalescent history will be complicated a bit by migrations between demes, but could be envisioned within a closely related inference framework.

Second, it could be interesting to use smoothing priors for \( N \) and \( S \) to ensure that these two functions of time do not show huge steps from a time period to the next. This will as well be the focus of future developments of the method, that could build on related work in phylodynamics (Karcher et al. 2020; Parag et al. 2021). It would seem especially interesting for \( N \) to root smoothing priors on mechanistic assumptions of population dynamics, such as considered e.g. in other epidemiological models (Cori et al. 2013).

Last, the boundaries of the intervals on which \( N \) and \( S \) remain constant could be unknown, with or without
prior distribution. This would allow the timeline to be informed by the data, and could lead as well, in case it is stochastic and averaged over the posterior, to smoother \( N \) and \( S \) trajectories. This could build on related work in a phyldynamic context under a birth-death model (Stadler 2011).

Further theoretical work is also required before this kind of model can be applied to larger – already available – datasets. Indeed, while the initial hope of bringing back into fashion the infinite allele model was to be able to process very large genetic datasets, it is in the end not realistic to use the current implementation of the method with more than \( \sim 10^4 \) sequences. At least four directions can be envisioned to improve on this objective in the future: (i) trying to optimize the sampling step of the coalescent history, which is the current bottleneck of the computation, so that it scales closer to \( O(B) \) than \( O(B^2) \), possibly using well designed approximations; (ii) turning to a maximum likelihood framework, possibly relying on an expectation-maximization algorithm to either completely bypass the need for posterior sampling through MCMC, or at least use it to drastically speed-up the burn-in of the MCMC; (iii) turning to a variational Bayes method properly designed to sample the posterior in our model; and (iv) optimizing the implementation of Algorithms 1 and 2, where the most straightforward idea would be to parallelize MCMC chains for example.

**When is the use of a simplified mutation process pertinent?**  A natural question arises when thinking about the difference between this method and currently used coalescent-based method in phyldynamics, namely *How do these compare in terms of statistical power?* Or to rephrase it in more technical terms: *what signal do we lose by forgetting about the coalescent history above the first mutation?*  The answer will likely depend on the mutation rate and on the temporal scale that one is interested in studying.

In the limit of very high mutation rate compared to the temporal scale under study, only singletons are observed, bearing no useful signal, while in the limit of very low mutation rate, only one allele is observed, again bearing no useful signal to infer \( N \) and \( S \).

In between, there is a setting with an intermediate mutation rate, such that alleles extend for some time across the focal time-frame, providing signal to reconstruct \( N \) and \( S \). However, even in this optimal setting for the method, data is being discarded so that all signal on the internal branches linking different alleles is lost as compared to coalescent-based methods integrating over the full unknown tree. When does the trade-off between computation time and precision turn in favor of using a simpler infinite alleles model? Quantifying this more precisely on simulations would be a valuable contribution.

Moreover, when one is not interested in estimating \( \mu \), the allele partition could be chosen so as to tend to an optimal setting as described above. In principle, the allele partition of the set of sequences could be obtained by applying any another equivalence relation, e.g. *being similar only on a given subsequence*, or *having a similar amino-acid sequence*. These could be used to decrease the number of alleles in the dataset.

In between the two above-mentioned extremes of (i) using an infinite alleles model or (ii) using a finite site model with a substitution model, lies also the opportunity to revive another assumption from population genetics, namely the *infinite sites model*. Under this model, each mutation hits a new site along the sequence, and thus a more precise phylogenetic history between sequences can be reconstructed. An *Importance Sampler* algorithm has been proposed by Stephens and Donnelly (2000) for simulating the past history under a coalescent with an infinite sites model. A more thorough comparison of this inference framework against another inference method relying on the infinite sites model could as well be a relevant contribution to the field.

Finally, this manuscript also opens the way to develop better approximations aiming at taking into accounts more sequences in scenarios with high numbers of duplicates (Boskova and Stadler 2020). This line of research could benefit from the joint use of different mutation models clearly distinguished in distinct parts of the evolutionary
tree of sequences, while still relying on a unique underlying population dynamics model, such as e.g. a coalescent model with discrete population size shifts.

**Conclusion** Bringing back into fashion old population genetics simplifications of the mutation process and incorporating them into modern statistical frameworks could play a key role in better surveying and understanding population demographics and structure from molecular data. I hope that this work will participate in a current trend towards adapting computer-intensive phylogenetics methods for use with datasets characterized by low genetic diversity such as the current SARS-CoV-2 outbreak.

**Acknowledgments**

This work was supported by an ETH Zürich Postdoctoral Fellowship. I thank the cEvo group at the D-BSSE department for interesting discussions on phylogenetics methods.

**References**

Barido-Sottani, J., T. G. Vaughan, and T. Stadler. 2020. A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. Systematic Biology 69:973–986.

Beaulieu, J. M. and B. C. O’Meara. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology 65:583–601.

Beerli, P. and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Boskova, V. and T. Stadler. 2020. PIQME: Bayesian phylogenetic method for analysis of large data sets with duplicate sequences. Molecular Biology and Evolution 37:3061–3075.

Cori, A., N. M. Ferguson, C. Fraser, and S. Cauchemez. 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology 178:1505–1512.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Durrett, R. 2008. Probability models for DNA sequence evolution. Springer.

Etienne, R. S., B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society B: Biological Sciences 279:1300–1309.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112.

Ewing, G., G. Nicholls, and A. Rodrigo. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 168:2407–2420.

Gonzalez-Reiche, A. S., M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, et al. 2020. Introductions and early spread of SARS-CoV-2 in the New York city area. Science 369:297–301.
Gupta, A., M. Manceau, T. Vaughan, M. Khammash, and T. Stadler. 2020. The probability distribution of the reconstructed phylogenetic tree with occurrence data. Journal of Theoretical Biology 488:110115.

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

Hein, J., M. Schierup, and C. Wiuf. 2004. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford university press.

Hörmann, W. and J. Leydold. 2014. Generating generalized inverse gaussian random variates. Statistics and Computing 24:547–557.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201.

Karcher, M. D., L. M. Carvalho, M. A. Suchard, G. Dudas, and V. N. Minin. 2020. Estimating effective population size changes from preferentially sampled genetic sequences. PLoS Computational Biology 16:-1007774.

Kendall, D. G. 1948. On the generalized ‘birth-and-death’ process. Ann. Math. Stat. 19:1–15.

Kim, K., R. Omori, and K. Ito. 2017. Inferring epidemiological dynamics of infectious diseases using Tajima’s D statistic on nucleotide sequences of pathogens. Epidemics 21:21–29.

Kingman, J. F. C. 1982. The coalescent. Stochastic processes and their applications 13:235–248.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429–434.

Lanave, C., G. Preparata, C. Sacone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86–93.

Lartillot, N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. Journal of Computational Biology 13:1701–1722.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Lemey, P., S. Hong, V. Hill, G. Baele, C. Poletto, V. Colizza, Á. O’Toole, J. T. McCrone, K. G. Andersen, M. Worobey, M. I. Nelson, A. Rambaut, and M. A. Suchard. 2020. Accommodating individual travel history, global mobility, and unsampled diversity in phylogeography: a SARS-CoV-2 case study. bioRxiv.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669–2680.

Leventhal, G. E., H. F. Günthard, S. Bonhoeffer, and T. Stadler. 2013. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Molecular Biology and Evolution 31:6–17.

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, et al. 2020. Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. The New England Journal of Medicine 382:1199–1207.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a binary character’s effect on speciation and extinction. Systematic Biology 56:701–710.
Maliet, O., F. Hartig, and H. Morlon. 2019. A model with many small shifts for estimating species-specific diversification rates. Nature Ecology & Evolution 3:1086–1092.

Manceau, M., A. Gupta, T. Vaughan, and T. Stadler. 2021. The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data. Journal of Theoretical Biology 509:110400.

McDonald, J. H. and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011. Reconciling molecular phylogenies with the fossil record. P. Natl. Acad. Sci. USA 108:16327–16332.

Müller, N. F., D. A. Rasmussen, and T. Stadler. 2017. The structured coalescent and its approximations. Molecular Biology and Evolution 34:2970–2981.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344:305–311.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. 2008. Genes mirror geography within Europe. Nature 456:98–101.

Parag, K. V., L. du Plessis, and O. G. Pybus. 2020. Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. Molecular Biology and Evolution 37:2414–2429.

Parag, K. V., O. G. Pybus, and C.-H. Wu. 2021. Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? Systematic Biology.

Pybus, O. G., A. J. Drummond, T. Nakano, B. H. Robertson, and A. Rambaut. 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Molecular Biology and Evolution 20:381–387.

Pybus, O. G. and P. H. Harvey. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. P. Roy. Soc. Lond. B. Bio. 267:2267–2272.

Rasmussen, D. A., O. Ratmann, and K. Koelle. 2011. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Computational Biology 7:1–1002136.

Skoglund, P., P. Sjödin, T. Skoglund, M. Lascoux, and M. Jakobsson. 2014. Investigating population history using temporal genetic differentiation. Molecular Biology and Evolution 31:2516–2527.

Stadler, T. 2010. Sampling-through-time in birth–death trees. Journal of Theoretical Biology 267:396–404.

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. P. Natl. Acad. Sci. USA 108:6187–6192.

Stadler, T., D. Kühnert, S. Bonhoeffer, and A. J. Drummond. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110:228–233.

Stephens, M. and P. Donnelly. 2000. Inference in molecular population genetics. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62:605–635.
Talts, S., M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. 2018. Validating Bayesian inference algorithms with simulation-based calibration. arXiv:1804.06788 [stat] ArXiv: 1804.06788.

Tavaré, S. 2004. Part I: Ancestral inference in population genetics. Pages 1–188 in Lectures on probability theory and statistics. Springer.

Vaughan, T. G., D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond. 2014. Efficient Bayesian inference under the structured coalescent. Bioinformatics 30:2272–2279.

Vaughan, T. G., G. E. Leventhal, D. A. Rasmussen, A. J. Drummond, D. Welch, and T. Stadler. 2019. Estimating epidemic incidence and prevalence from genomic data. Molecular Biology and Evolution 36:1804–1816.

Vaughan, T. G., J. Sciré, S. A. Nadeau, and T. Stadler. 2020. Estimates of outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data. medRxiv.

Volz, E. M., S. L. K. Pond, M. J. Ward, A. J. L. Brown, and S. D. Frost. 2009. Phylodynamics of infectious disease epidemics. Genetics 183:1421–1430.
A Details on the numerical method

Pipeline for creating the allele partition

Creating the allele partition is a central step in the method, for this will be the raw data taken as input by the Gibbs sampler. To build it, I relied on the following pipeline:

1. cut the master reference sequence from position 250 to 29700, and put this one as the first element of a list of reference sequences.
2. initialize an empty list of alleles.
3. Then iterate, for each other sequence, the following steps,
   (a) look for a pattern close to the 30 first nucleotides of the master reference in to the first 500 nucleotides of each sequence, to get the beginning of the window, and discard the very few sequences where the beginning could not be found.
   (b) keep as the focal sequence the 29450 nucleotides following the beginning of the window.
   (c) record the collection of snps differing in the focal sequence as compared to the reference sequence.
      • if it is smaller than 100 snps long, we consider it to be the compressed representation of the sequence against this reference.
      • it not, compare to the next reference sequence in the list.
   (d) if no reference sequence is less than 100 snps different than the sequence, it is likely to have a feature of its own (typically, a gap). Add this sequence to the list of reference sequences.
   (e) Look for the same compressed representation of the sequence in the list of alleles and add the sampling date of this sequence to the list of sampling dates of the allele.

This pipeline is implemented in the script raw_to_datasets.ml, available as part of the code associated to this article.

MCMC output analysis

MCMC samples output by the Ocaml code are then analyzed using scripts written in the R programming language. Post-processing steps rely in particular on the following packages to work,

• the very versatile ggplot2 and cowplot to produce figures,
• LaplacesDemon to compute ESS values using traces of scalar values,
• forecast to compute ACF or PACF using traces of scalar values,

The R post-processing scripts are also available as part of the code released along this article.
Generation of random variables

Random variables distributed according to a GIG distribution are sampled using a personal implementation of Hörmann and Leydold (2014)’s algorithm in the programming language Ocaml. It is naive translation of the very handy R package GIGrvg by the same authors.

B Case count data in the four countries

Case count numbers in the four countries of interest have been downloaded from the WHO website on the 16th of June 2021. I plot here the number of samples, number of sequences, and sampling fraction through time over the time frame of the first wave of the outbreak, for comparison with the order of magnitudes of $N$ and $S$ presented in the main text.

Note here that the sequencing fraction plotted below corresponds to the number of sequences divided by the sum of the number of sequences and number of samples. Indeed, the number of sequences in the early time period is sometimes higher than the number of samples, and I thus considered that the sequences were not included in the case count.

![Figure 7: Information on the number of infected, and the number of sequences, sampled through time in CH, DE, FR, and IT.](image)

(region
- CH
- DE
- FR
- IT)

(region
- CH
- DE
- FR
- IT)
C Gibbs sampler code documentation

This Section aims at making the code released along this article more easily comprehensible, by shortly presenting in plain text the strategy to store different data, together with key functions to manipulate the data.

Format of key quantities

By default, all lists are ordered in time ascending order (that is, first element on top of the list at present time 0).

listS or listN list of ordered \((t_j, S_j)\), where \(S_j\) is valid on \([t_j, t_{j+1})\).

an allele or a lineage an ordered list of either \((b_i)\) belonging to the same allele/lineage, if the individuals are not numbered, or an ordered list of \((n_i)\) if the individuals have been numbered.

alleles or lineages a list of (ordered lists of \((b_i)\)) or a list of (ordered lists of \((n_i)\)) depending again on whether individuals are numbered or not.

array_individuals an array of information about individuals, with element number \(i\) referring to individual number \(i\). Each element is a tuple \((b_i, h_i, o_i, a_i)\) where \(b_i\) is the birth time, \(d_i\) is the death time, \(o_i\) is the number of the individual into which this individual coalesces (or \(o_i = i\) if there is a mutation), and \(a_i\) is the ID number of the allele this individual belongs to.

array_alleles an array of information about alleles. Each element number \(a\) is the list of ID numbers of all individuals belonging to allele \(a\).

samp_history an ordered list of \((b_i, i)\) where \(i\) is the ID number of the newly born individual at time \(b_i\).

coal_history an ordered list of \((h_i, i, o_i)\) where \(i\) is the ID number of the individual dying at time \(h_i\) by coalescence into lineage \(o_i\) (or by a mutation if \(o_i = i\)).

all_events a list of all events ordered in time ascending order, with lots of other interesting quantities attached to the interval between this event and the following, such as : the number of lineages alive, the current values of \(N\), the current value of \(S\), the total rate at which an event happens on this interval for a given individual, etc...

intervals the ordered in DESCENDING order list of \((t_s, t_e, k, k_a^{<i}, \tau, w)\) where \(t_s\) is the time of start, \(t_e\) is the end time, \(k\) is the total number of lineages alive on the interval, \(k_a^{<i}\) is the number of individuals from allele \(a\), to the left of individual \(i\), being alive on the interval, \(\tau\) is the total rate at which death happens and \(w\) is the weight of the interval, i.e. the probability that the death event of the focal individual falls in this interval, conditioned on \(A\). Note that the reason it is in descending order is that it is built by reading the list of all_events in ascending order, and the next steps consisting in drawing the interval does not require any specific order.

Roles of some key functions

Here is now an overview of some of the key operations we need to perform for Gibbs sampling steps.

simulation if needed to get samp_history first using sim_sampling_events.

And then to get alleles using sim_coal.
**pre-processing** we build the `array_individuals` and `array_alleles` from the `alleles` and `samp_history`.

**Gibbs sampling** We need to consider in turn the following operations,

- Updating the death time of individual number \(i\) using `update_past_coal`:
  requires to first build the weighted intervals with `get_intervals`,
  before drawing an interval with `find_back_interval`,
  simulating the death time within the interval and finally erasing the old time and inserting the new at a correct place in `all_events` using `replace_coal`.

- Updating the \((N_j)\) values using `update_listN`, which explores the list of `all_events`, and modifies in place the \(N_j\) values.

- Updating the \((S_j)\) values using `update_listS`, working similarly as above.

- Updating \(\mu\) using `update_mu`

**D Acknowledgements for data collection**

I am grateful to the following list of authors, who have contributed to the collection of SARS-CoV-2 sequences that I downloaded on GISAID. In the four tables below, you will find the authors that took part in the collection effort in, respectively, Switzerland, Germany, France and Italy over the time period I have been interested in.
We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based.

All Submitters of data may be contacted directly via www.gisaid.org

Authors are sorted alphabetically.

**Accession ID**  
EPI_ISL_1129072, EPI_ISL_1129081, EPI_ISL_1129082, EPI_ISL_1129083, EPI_ISL_1129084, EPI_ISL_1129085, EPI_ISL_1129086, EPI_ISL_1129087, EPI_ISL_1130034, EPI_ISL_1130035, EPI_ISL_1130036, EPI_ISL_1130037, EPI_ISL_1130038, EPI_ISL_1130039, EPI_ISL_1130040, EPI_ISL_1130041, EPI_ISL_1130042

**Originating Laboratory**  
University Hospital Basel, Clinical Virology

**Submiting Laboratory**  
University Hospital Basel, Clinical Virology

**Authors**  
Trestan Pillonel, Damien Jacot, Sébastien Aeby, Gilbert Greub, Claire Bertelli
Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich
Institute of Medical Virology, University of Zurich
Institute for Infectious Diseases, University of Bern
Department of Infectious Diseases, Cantonal Hospital Baden
Division of Infectious Diseases, University Hospital Zürich
Institute of Medical Virology, University of Zurich
Hospital Schwyz
Institute of Medical Virology, University of Zurich
Division of Infectious Diseases, University Hospital Zürich
Department of Infectious Diseases, Cantonal Hospital Baden
Division of Infectious Diseases, University Hospital Zürich
Hospital Schenecty
Division of Infectious Diseases, University Hospital Zürich
see above
University Hospital Basel, Clinical Virology
University Hospital Basel, Clinical Bacteriology
see above
Institute for Infectious Diseases, University of Bern
Institute for Infectious Diseases, University of Bern
see above
| Accession ID | Originating Laboratory | Submitting Laboratory | Authors |
|-------------|------------------------|-----------------------|---------|
| EPI_ISL_114086 | Center of Medical Microbiology, Virology, and Hospital, Hygiene, University of Duesseldorf | Robert Koch Institute | unknown |
| EPI_ISL_114088 | Center of Medical Microbiology, Virology, and Hospital, Hygiene, University of Duesseldorf | Robert Koch Institute | unknown |
| EPI_ISL_114089 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_457750 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_459962 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_452218 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_452219 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_451945 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_451940 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450200 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450203 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450198 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450199 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450192 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450193 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450194 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450195 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450196 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450197 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450198 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450199 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450200 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450203 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450204 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450205 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450206 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450207 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450208 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450209 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450210 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450211 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450212 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450213 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450214 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450215 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450216 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450217 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450218 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450219 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450220 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450221 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450222 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450223 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450224 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450225 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450226 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450227 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450228 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450229 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450230 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450231 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450232 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450233 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450234 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450235 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450236 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450237 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450238 | Robert Koch Institute | Robert Koch Institute | unknown |
| EPI_ISL_450239 | Robert Koch Institute | Robert Koch Institute | unknown |

We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based.

All Submitters of data may be contacted directly via gisaid.org

Authors are sorted alphabetically.
Bonn-Cologne, Bonn, Germany

Marek Korencak et al

University Medical Center Hamburg Eppendorf

UK Tübingen, Medical Microbiology

Charité Universitätsmedizin Berlin, Institut für Virologie/Labor Billo

Charité Universitätsmedizin Berlin, Institut für Virologie/Virology

Virology

Alexis Robitaille, Thomas Günther, Johannes Kindlisch, Marte Andechs, Nicole Fischer, Aden Grundolf

University of Freiburg, Medical Microbiology, Ulm University Medical Center, Kirchhoff Lab / Ulm University Medical Center, Virology

Christian Drosten

Jonas Fuchs, Lisa Kern, Sandra Reuter, Hejo Grundmann, Marcus Fanning

Philipp Schoch, David Hather, Kevin J. Kelleher, Stephanie Schmaljohn, Thomas Ruge, Christian H. Morschhäuser, Michael G. Stöver, Christian Drosten

Christian Drosten

Jonas Fuchs, Lisa Kern, Sandra Reuter, Hejo Grundmann, Marcus Fanning

David Hather, Kevin J. Kelleher, Stephanie Schmaljohn, Thomas Ruge, Christian H. Morschhäuser, Michael G. Stöver, Christian Drosten

Martin Aepfelbacher, Nicole Fischer, Adam Grundhoff

Christian Drosten

Jonas Fuchs, Lisa Kern, Sandra Reuter, Hejo Grundmann, Marcus Fanning

David Hather, Kevin J. Kelleher, Stephanie Schmaljohn, Thomas Ruge, Christian H. Morschhäuser, Michael G. Stöver, Christian Drosten

Christian Drosten

Jonas Fuchs, Lisa Kern, Sandra Reuter, Hejo Grundmann, Marcus Fanning

David Hather, Kevin J. Kelleher, Stephanie Schmaljohn, Thomas Ruge, Christian H. Morschhäuser, Michael G. Stöver, Christian Drosten
We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens, as well as the

All Submitters of data may be contacted directly via www.gisaid.org

Authors are sorted alphabetically.
Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center

Virology Laboratory, Scientific Department, Army Medical Center
