RESEARCH ARTICLE

adaPop: Bayesian inference of dependent population dynamics in coalescent models

Lorenzo Cappello[1☯], **Jaehee Kim**[2☯], **Julia A. Palacios**[3*]

1 Departments of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain, **2** Department of Computational Biology, Cornell University, Ithaca, New York, United States of America, **3** Departments of Statistics and Biomedical Data Science, Stanford University, Stanford, California, United States of America

☯ These authors contributed equally to this work.
* juliapr@stanford.edu

Citation: Cappello L, Kim J, Palacios JA (2023) adaPop: Bayesian inference of dependent population dynamics in coalescent models. PLoS Comput Biol 19(3): e1010897. https://doi.org/10.1371/journal.pcbi.1010897

Editor: Jennifer A. Flegg, The University of Melbourne Melbourne School of Psychological Sciences, AUSTRALIA

Received: July 27, 2022

Accepted: January 25, 2023

Published: March 20, 2023

Data Availability Statement: The implementation of our methods adaPop is available in the R package adapref (https://github.com/lorenzocapp/adapref). The codes to reproduce the simulation study and real data analysis are available at https://github.com/lorenzocapp/adapop_numexp.

Funding: JAP was supported by National Institutes of Health grant R01-GM-131404 and NSF Career Award #2143242. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The coalescent is a powerful statistical framework that allows us to infer past population dynamics leveraging the ancestral relationships reconstructed from sampled molecular sequence data. In many biomedical applications, such as in the study of infectious diseases, cell development, and tumorgenesis, several distinct populations share evolutionary history and therefore become dependent. The inference of such dependence is a highly important, yet a challenging problem. With advances in sequencing technologies, we are well positioned to exploit the wealth of high-resolution biological data for tackling this problem. Here, we present `adaPop`, a probabilistic model to estimate past population dynamics of dependent populations and to quantify their degree of dependence. An essential feature of our approach is the ability to track the time-varying association between the populations while making minimal assumptions on their functional shapes via Markov random field priors. We provide nonparametric estimators, extensions of our base model that integrate multiple data sources, and fast scalable inference algorithms. We test our method using simulated data under various dependent population histories and demonstrate the utility of our model in shedding light on evolutionary histories of different variants of SARS-CoV-2.

Author summary

Genomic data provide information about evolutionary dynamics—such as an evolving epidemic and tumorgenesis—that is difficult to infer from other source of data. One of the main computational challenges in inferring past population histories is to jointly model dependent subpopulations and correctly quantifying their dependencies over time. When distinct subpopulations have common ancestry in the past and evolve under shared environmental pressure, their population dynamics become dependent. In this work, we propose an efficient inference method for studying dependent population dynamics from genetic data in the coalescent framework: an approach that considers the stochastic process of the "coalescence" of genealogical lineages traveling back in time to explain the statistical properties of a sample's genetic variation. We also extend our framework to jointly model the ancestral and sampling processes incorporating sampling times as an additional

Competing interests: The authors have declared that no competing interests exist.

source of information. We validate our methods via extensive simulations and demonstrate that our methods provide new insights into the evolutionary dynamics of SARS-CoV-2 novel variants.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Bayesian inference of past population sizes from genetic data is an important task in molecular epidemiology of infectious diseases and other biomedical disciplines [1–4]. While many computational and methodological advances have been developed in the last 20 years, there are still many challenges in using these models on real data applications (see [5, 6] for recent reviews).

One of the important limitations of most current methods is their lack of flexibility in modeling the dependency of populations' effective population sizes. Current models typically assume either single population size dynamics or a structured population undergoing migration. In particular, when modeling structured populations, these models resort to simplistic assumptions on the population size dynamics in order to gain computational tractability and parameter identifiability [7, 8]. However, often there are situations where the population is neither one large unit nor completely divided into isolated subpopulations. Different subpopulations may share the same environment and partial ancestry some time in the past, and therefore their population dynamics are dependent. For example, in the study of infectious diseases, all variants of a virus may share the same environment and local non-pharmaceutical interventions; hence their population dynamics could be similarly affected by these external interventions. In tumorgenesis, the cancer cell population within an individual undergoes clonal expansions in the tumor microenvironment, often resulting in a mixture of genotypically and phenotypically heterogenous cell subpopulations. Identifying and quantifying such expansions is crucial for timely detection and personalized oncology for cancer [9]. Despite its importance, to the best of our knowledge, no realistic methods exist for jointly modeling and studying the trajectories of dependent population size trajectories.

In this work, we propose a nonparametric method for inferring dependent past population dynamics of subpopulations and for estimating the relative difference in their population size trajectories over time. The proposed method bypasses the problems inherent in modeling complex dependent population dynamics by *a priori* modeling the dependency of population sizes via a nonparametric prior. Our method not only provides a measure of this dependence, but also increases estimates' accuracy and generates narrower credible bands. Essentially, this happens because we can employ more data to estimate the parameter of interest. In addition, our approach incorporates other types of data informative of the parameter of interest such as temporal sampling information of molecular sequences.

Our contributions can be summarized as follows.

- We propose a nonparametric Bayesian framework for inferring dependent population dynamics from genetic data in the coalescent framework. The model makes minimal assumptions on the functional form of the population trajectories and their dependency. Despite its flexibility, we prove that model parameters are identifiable.

- We extend our framework to jointly model the ancestral and sampling processes incorporating sampling times as an additional source of information. We empirically validate the performance of our methods and show the ability of the sampling-aware methods in reducing bias and improving estimation.
- We demonstrate the utility of our methods in providing new insights into the evolutionary dynamics of SARS-CoV-2 novel variants.

Background

Coalescent model

The coalescent [10] is a popular prior model on genealogies. The genealogy is a timed and rooted binary tree that represents the ancestry of a sample of n individuals from a population. We assume that g is a discrete ranked and labeled tree topology with n leaves, $(n_\ell)_{\ell=1:m}$ samples are sequentially collected at m sampling times denoted by $\mathbf{s} = (s_\ell)_{\ell=1:m}$, with $s_1 = 0$, $s_{j-1} < s_j$ for $j = 2, \ldots, m$, and $n = \sum_{j=1}^{m} n_j$ is the total number of samples. In the genealogy, pairs of lineages merge backwards in time into a common ancestor at coalescent times denoted by $\mathbf{t} = (t_n, \ldots, t_2)$ (Fig 1). The rate at which pairs of lineages coalesce depends on the number of lineages and the effective population size (EPS) denoted by $(N_e(t))_{t \geq 0} \coloneqq N_e$. Under this model, the density of a genealogy $\mathbf{g} = (g, \mathbf{t}, \mathbf{s}, \mathbf{n})$ is:

$$p(\mathbf{g} \mid N_e) = \exp\left(-\int_0^\infty \frac{C(t)}{N_e(t)}\,dt\right)\prod_{k=2}^{n}\frac{1}{N_e(t_k)}, \tag{1}$$

where $C(t) = \binom{A(t)}{2}$ is the coalescent factor—a combinatorial factor of the number of extant lineages $A(t) = \sum_{i=1}^{m} n_i \mathbb{1}(s_i < t) - \sum_{k=2}^{n} \mathbb{1}(t_k < t)$. The EPS is generally interpreted as a relative measure of genetic diversity [11].

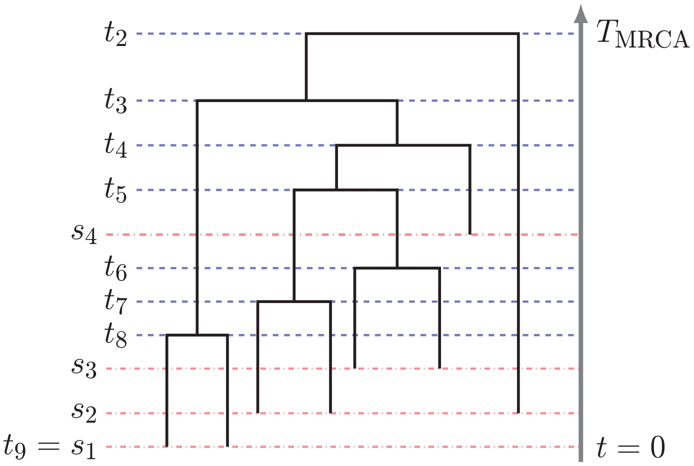

Fig 1. Example of a genealogy with sequential sampling. $s_1, \ldots, s_4$ and $t_8, \ldots, t_2$ indicate sampling times (red dotted lines) and coalescent times (blue dotted lines), respectively. The time increases backwards in time starting with $s_1 = t_9 = 0$ as the present time. T_{MRCA} indicates the time to the most recent common ancestor at the root.

https://doi.org/10.1371/journal.pcbi.1010897.g001

Bayesian inference of EPS

We start by assuming that a given genealogy is available to us. Bayesian inference of the EPS then targets

$$P(N_e, \boldsymbol{\tau} \mid \mathbf{g}) \propto P(\mathbf{g} \mid N_e)P(N_e \mid \boldsymbol{\tau})P(\boldsymbol{\tau}), \tag{2}$$

where $P(N_e|\boldsymbol{\tau})$ is a prior distribution on N_e that depends on a vector of hyperparameters $\boldsymbol{\tau}$.

A common choice for a prior on N_e is to assume an underlying finite-dimensional parametric structure. In particular, this choice makes the calculation of the integral in Eq (1) computationally tractable. A popular strategy is to use a regular grid of $M + 1$ points $(k_i)_{1:M+1}$ and assume that N_e is well approximated by

$$N_e(t) = \sum_{i=1}^{M+1} \exp(\theta_i)\mathbb{1}(t \in (k_i, k_{i+1})), \tag{3}$$

a piece-wise constant function with M change points [1, 12]. In Eq (3), we model N_e in log-scale, however this is not strictly necessary. Many approaches place a Gaussian Markov random field (GMRF) prior on $\boldsymbol{\theta}$, for example [13–16]. An alternative to the piece-wise constant assumption is to use Gaussian process priors. However, the posterior distribution becomes doubly intractable because the likelihood function depends on an infinite-dimensional integral over Gaussian processes. In [17], the authors augmented the posterior distribution with auxiliary variables via thinning of Poisson processes [18] in order to gain tractability.

The advantage of the GP prior is that one does not need to specify a grid. This comes at the additional computational costs mentioned above. GMRFs priors require selecting a grid; however, current formulations (e.g., [16]) leave ample flexibility in the choice of the number of parameters M and grid breakpoints while remaining computationally extremely tractable. This is the case because neither the grid cell boundaries $(k_i)_{1:M+1}$ nor M depend on the data $\mathbf{g} = (g, \mathbf{t}, \mathbf{s}, \mathbf{n})$ or the model parameters $N_e(t)$. By increasing M, one effectively enriches the family of functions supported by the prior, which is the motivation underlying the use of a GP. We refer to [16] for guidance on the choice of M.

There are exact and approximate algorithms to sample from Eq (2). Standard software packages [3, 19] employ Markov chain Monte Carlo (MCMC) methods with carefully designed transition kernels. Recent algorithmic advances employ Hamiltonian Monte Carlo (HMC). [16] implemented an algorithm to sample from Eq (2) under several Markov random fields priors on STAN [20], and Lan et al. [21] employed split HMC [22]. Among the approximate methods, inference can be efficiently done with Integrated Nested Laplace Approximation (INLA) [23]. In [24], authors use INLA under a GMRF prior on N_e, showing that the approximate posterior is remarkably similar to that obtained by MCMC-based algorithms.

Preferential sampling

The standard coalescent model implicitly assumes that the sampling times are either fixed or functionally independent of the underlying population dynamics. This assumption is in stark contrast with birth-death-sampling models, where one needs to specify a sampling process along with the evolutionary model [25]. However, in many applications, such as in infectious diseases, the sampling frequency is often highly correlated with EPS: more samples are sequenced when the EPS is larger. This situation, known as preferential sampling in spatial statistics [26], allows us to model sampling frequency information in order to improve inference about EPS, reducing estimation bias and improving the accuracy of model parameter inference. A parametric preferential sampling model was first introduced in coalescent inference by

[25] and later extended to the nonparametric setting [27–30]. The probabilistic dependency of sampling time distribution on population dynamics can be modeled as an inhomogeneous Poisson point process (iPPP) with a rate $\lambda(t)$ that depends on EPS. Although in practice, sampling events occur in bulk, we assume that samples arrive at an instantaneous rate within a time interval. Hence, we approximate the likelihood of sampling events by the counting process over a fixed grid.

We adopt the adaptive preferential sampling framework [30] that employs a flexible approach for modeling the time-varying dependency between $N_e(t)$ and $\lambda(t)$: $\lambda(t) = \zeta(t)N_e(t)$, where both $\zeta(t)$ and $N_e(t)$ are unknown continuous functions with GMRF priors.

Methods

We propose a flexible and scalable framework for modeling the genealogies of two samples (not necessarily of the same size) from two populations with dependent population size dynamics. The goals are (i) to estimate the EPSs of the two populations, (ii) to account for the dependency between them, and (iii) to quantify estimation uncertainty. The implementation of our methods `adaPop` is available in the `R` package `adapref` (https://github.com/lorenzocapp/adapref).

Coalescent for dependent population size dynamics

Let $\mathbf{g}^A = (g^A, \mathbf{t}^A, \mathbf{s}^A, \mathbf{n}^A)$ and $\mathbf{g}^B = (g^B, \mathbf{t}^B, \mathbf{s}^B, \mathbf{n}^B)$ be the genealogies of samples collected from populations A and B respectively, and let N_e^A and N_e^B denote their corresponding EPSs. Here, $(\mathbf{s}^A, \mathbf{s}^B)$ are the vectors of sampling times, and $(\mathbf{n}^A, \mathbf{n}^B)$ are the corresponding vectors of number of samples collected. Standard coalescent-based inference methodologies ignore any association between the underlying population processes of the two populations when approximating posterior distributions $P(N_e^A|\mathbf{g}^A)$ and $P(N_e^B|\mathbf{g}^B)$.

The advantages of directly modeling the dependence are twofold. First, we get a direct measure of the association that can have a direct interpretation in scientific studies. Second, our hierarchical model should estimate N_e more accurately because we model N_e as a shared parameter, hence we borrow information from the two samples, which is a standard advantage of a hierarchical Bayesian model [31].

We model the association between the two population size trajectories, which can change over time, with a time-varying parameter linking N_e^A and N_e^B:

$$\begin{aligned}
\mathbf{g}^A \mid N_e, \mathbf{s}^A, \mathbf{n}^A &\sim \text{Coalescent}\,(N_e),\\
\mathbf{g}^B \mid N_e, \gamma, \mathbf{s}^B, \mathbf{n}^B &\sim \text{Coalescent}\,(\gamma\, N_e),\\
\log N_e &\sim \text{GMRF}(\tau_1),\\
\tau_1|a, b &\sim \text{Gamma}(a, b),\\
\log \gamma|\tau_2 &\sim \text{GMRF}(\tau_2),\\
\tau_2|a, b &\sim \text{Gamma}(a, b).
\end{aligned} \tag{4}$$

Here, $\gamma := (\gamma(t))_{t \geq 0}$ is the time-varying coefficient that describes how the association between the two population processes changes over time, leading to $N_e^A = N_e$ and $N_e^B = \gamma N_e$. The interpretation of γ provides information on the association between two populations. For example, a growing trend in γ signals the existence of an association between two EPSs: N_e^B is growing faster than N_e^A. However, it does not necessarily imply a positive association because, for example, if N_e^B were growing, N_e^A could be either growing at a slower rate than N_e^B or be decreasing and still have an increasing γ.

Throughout the section, we employ GMRF priors with precision τ_1 and τ_2 on N_e and γ; however, the framework is flexible to any prior distribution. It is possible to go fully nonparametric employing a GP [17], or a different kind of MRFs, for example, the Horseshoe MRFs [32]. The number of parameters of the two GMRFs tunes how free the dependence is allowed to vary between the two populations. In the numerical illustrations, we will employ GMRFs modeling first order dependencies.

Our model in Eq (4) is "asymmetric", in the sense that the baseline population EPS is multiplied by the time-varying coefficient γ to define the EPS of a new population. The choice is motivated by the actual scientific question we are examining, in which a new population develops from an existing one.

We will compare our proposal with a simpler parametric model suggested in [5]. Here, the population dependence is modeled through two time-independent scalar parameters α and β:

$$N_e^A = N_e \text{ and } N_e^B = \alpha (N_e)^{\beta}. \tag{5}$$

However, the strict parametric dependence enforced in the model increases the risk of model misspecification by not allowing changes in the association of the two population processes. For example, the model support excludes a time shift when $N_e^B(t) = N_e^A(t+s)$ for $s > 0$. A consequence of this potential model misspecification is the biased estimation of N_e^B, α, and β; we will provide numerical proofs of this claim in Results section.

Preferential sampling and dependent population size dynamics

A common approach for studying the relative growth between two population dynamics is to model how the sampling frequency of molecular sequences changes over time in the two populations. This is frequently done by fitting logistic growth models to the sampling dates only [33, 34]. The probabilistic models discussed in the previous section take a different stance and employ molecular data to reconstruct the genealogies which in turn are used to infer the population processes jointly. Here, we extend the probabilistic models, either model Eqs (4) or (5), and model the genealogies jointly with the observed sampling frequencies from both samples as follows:

$$\begin{aligned}
\mathbf{s}^A|\zeta &\sim \text{iPPP}(\zeta\, N_e^A),\\
\mathbf{s}^B|\zeta &\sim \text{iPPP}(\zeta\, N_e^B),\\
\log\zeta &\sim \text{GMRF}(\tau_3),\\
\tau_3|a,b &\sim \text{Gamma}(a,b).
\end{aligned} \tag{6}$$

Eq (6) builds on the preferential sampling framework described in Background section: the sampling process is an iPPP whose rate is a function of both the EPS and a time-varying parameter ζ. Here, the sampling process of populations A and B will have distinct rates, $\lambda^A = \zeta\, N_e^A$ and $\lambda^B = \zeta\, N_e^B$. Although we assumed a shared ζ function, our implementation considers the possibility of one ζ function per population. We emphasize that the model is flexible to any choice of prior distributions on ζ.

Inference

We start describing the inference procedure for the parameters of the model in Eq (4). We employ the same discretization described in Background section: given a regular grid $(k_i)_{1:M+1}$, we assume that N_e is governed by parameters $\boldsymbol{\theta}$ through the map given in Eq (3); γ is governed by parameters $\boldsymbol{\xi} = (\xi_i)_{i\,=\,1:M'}$ such that $\gamma(t) = \exp \xi_i$ for $t \in (k_i, k_{i+1}]$. Let $\boldsymbol{\tau}$ be the vector of precision hyperparameters of the GMRFs. Following [24], we use INLA for obtaining marginal posterior medians and marginal 95% Bayesian credible intervals (BCI).

INLA does not approximate the full posterior $P(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{g}^A, \mathbf{g}^B)$; rather, it approximates the posterior marginals $P(\boldsymbol{\tau}|\mathbf{g}^A, \mathbf{g}^B)$, $(P(\theta_i|\mathbf{g}^A, \mathbf{g}^B))_{1:M}$, and $(P(\xi_i|\mathbf{g}^A, \mathbf{g}^B))_{1:M}$. The first step consists in computing

$$\widehat{P}(\boldsymbol{\tau}|\mathbf{g}^A, \mathbf{g}^B) \propto \frac{P(\xi, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)}{\widehat{P_G}(\xi, \boldsymbol{\theta}|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)}\Bigg|\begin{array}{l} \xi = \xi^*(\boldsymbol{\tau}), \\ \boldsymbol{\theta} = \boldsymbol{\theta}^*(\boldsymbol{\tau}) \end{array},$$

where the denominator is the Gaussian approximation to $P(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)$ obtained from a Taylor expansion around its modes $\boldsymbol{\theta}^*(\boldsymbol{\tau})$ and $\boldsymbol{\xi}^*(\boldsymbol{\tau})$ (the first Laplace approximation). The second step approximates the marginal posteriors of $P(\theta_i|\mathbf{g}^A, \mathbf{g}^B)$ and $P(\xi_i|\mathbf{g}^A, \mathbf{g}^B)$. For example $P(\theta_i|\mathbf{g}^A, \mathbf{g}^B)$ is approximated by

$$\widehat{P}(\theta_i|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B) \propto \frac{P(\xi, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)}{\widehat{P_{GG}}(\boldsymbol{\theta}_{-\boldsymbol{i}}, \xi|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)}\Bigg|\begin{array}{l} \boldsymbol{\theta}_{-\boldsymbol{i}} = \boldsymbol{\theta}^*_{-\boldsymbol{i}}, \\ \xi = \xi^*, \end{array}$$

where the denominator is a further Gaussian approximation of the corresponding conditional distribution. Now the Taylor expansion is centered at $(\boldsymbol{\theta}_{-\boldsymbol{i}}, \xi) = \mathrm{E}_G[\boldsymbol{\theta}_{-\boldsymbol{i}}, \boldsymbol{\xi}|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B]$, where the expected value is taken w.r.t. $\widehat{P_G}(\boldsymbol{\theta}, \xi|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)$. The subscript GG highlights the fact that two Gaussian approximations are employed to define P_{GG}. The last step involves integrating out the hyperparameters from $\widehat{P}(\theta_i|\boldsymbol{\tau}, \mathbf{g}^A, \mathbf{g}^B)$. This can be easily accomplished using $\widehat{P}(\boldsymbol{\tau}|\mathbf{g}^A, \mathbf{g}^B)$ (the nested Laplace approximation step).

Identifiability

Parameter identifiability is an essential property of models used in statistical learning. Roughly speaking, it refers to the theoretical possibility of uniquely estimating a parameter vector if an infinite amount of data is available [35–37]. Note that this is a property of the generative model, not of the estimator used.

For example, if $Y \sim Poisson(\lambda_1\lambda_2)$, then for a pair (λ'_1, λ'_2), any combination $\left(\frac{1}{c}\lambda'_1, c\lambda'_2\right)$ with $c > 0$ will be observationally equivalent. The multiplication of two parameters is a feature often leading to unidentifiability. Despite the fact that the models described in Eqs (4) and (5) include a product of parameters, we show that identifiability is not lost.

Since the parameters of models in Eqs (4) and (5) have a scientific interpretation, a lack of identifiability could hinder the validity of the scientific insights gained from using our methodology. There is a large literature showing that many models in evolutionary biology and ecology are not identifiable [5, 38]. In the coalescent literature, [39] shows that N_e is identifiable in the neutral and structured case. Our proposal does not fall in these two categories and a new result is required. Under the assumption that N_e and γ are piecewise-constant, *i.e.* $N_e = (N_{e,\,i})_{1:M}$ and $\gamma = (\gamma_i)_{1:M}$, we prove that the models introduced in this section possess this important property.

Proposition 1 (Identifiability of the nonparametric model, Eq (4)). *Let $\mathbf{g}^A$ be distributed as a coalescent with EPS $(N_{e,\,i})_{1:M}$, and $\mathbf{g}^B$ as a coalescent with EPS $(\gamma_i N_{e,\,i})_{1:M}$, $M \geq 1$, then the vector $(N_{e,1}, \ldots, N_{e,\,M}, \gamma_1, \ldots, \gamma_M)$ is identifiable.*

Proposition 2 (Identifiability of the parametric model, Eq (5)). *Let $\mathbf{g}^A$ be distributed as a coalescent with EPS $(N_{e,\,i})_{1:M}$, and $\mathbf{g}^B$ as a coalescent with EPS $(\alpha N_{e,i}^{\beta})_{1:M}$, if $M = 2$, the vector $(N_{e,1}, \ldots, N_{e,\,M}, \alpha, \beta)$ is identifiable.*

Proofs of Propositions 1 and 2 can be found at Section A in S1 Text. [39] proves identifiability of N_e for the standard coalescent employing results in [35], which consists in showing that the expected Fisher information is non-singular. We follow the same template.

Similar to [39], we require for Propositions 1 and 2 to hold at least one coalescent event within each interval in the grid because at least one data point is needed to have non-zero Fisher information. This is not specific to our setting; it is also true in the classical case with a single population [39] when using a skyline estimator [40]. However, this is a theoretical result that takes into account only the likelihood. Our method is quite different because we employ a Bayesian formulation with GMRFs priors that add "structure" to the estimation problem in the sense of enforcing smooth estimates. This is likely to alleviate the requirements of at least one coalescent event per grid interval as evidenced by the empirical success of GMRF [13–16]. The reason is that the extra information carried in the prior helps addressing the lack of observations in a given interval.

Extension to multiple populations

Our work is centered on a two-population model because one of our main applications targets the estimation of a relative advantage of a newly emerging viral variant over an existing variant. We further assume that the new variant originated from the standing variant. However, the framework can easily be extended to include multiple populations viewing the two-population model as a building block for a general genealogical model with multiple populations. In this generalization, the EPS of a child population is a function of the EPS of its parental population. This gives the two types of hierarchical structures displayed in Fig 2:

- *Nested populations.* Each effective population is a function of its immediate preceding one (Fig 2A). The baseline population "A" with EPS $N_e^A = N_e$ evolves into a second population "B" with EPS $N_e^B = \gamma_1 N_e$, which then in turn evolves into a population "C" with EPS $N_e^C = \gamma_1 \gamma_2 N_e$.

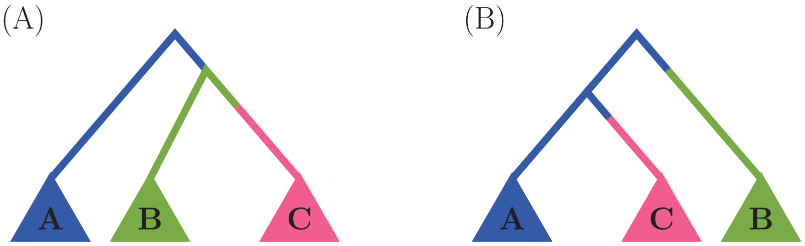

Fig 2. Modeling of multiple subpopulations with dependent population size dynamics. The trees represent large genealogies of many sequences from three different subpopulations labeled "A", "B" and "C" at the tips of the trees. Each lineage represents the subtrees of individuals whose rate of coalescence is dictated by the color of the branch. (A) Nested populations. The blue branch indicates coalescent events happen at rate that depends on $N_e^A = N_e$, green branch indicates a coalescent rate that depends on $N_e^B = \gamma_1 N_e$, and pink branch indicates a coalescent rate that depends on $N_e^C = \gamma_1 \gamma_2 N_e$. (B) Radial populations. The blue branch indicates a coalescent rate that depends on $N_e^A = N_e$, the green branch indicates a coalescent rate that depends on $N_e^B = \gamma_1 N_e$, and pink branch indicates a coalescent rate that depends on $N_e^C = \gamma_2 N_e$.

https://doi.org/10.1371/journal.pcbi.1010897.g002

- *Radial populations.* Multiple populations evolve from the baseline population "A" with EPS $N_e^A = N_e$ (Fig 2B). Then, population "B" will have EPS $N_e^B = \gamma_1 N_e$, and population "C" with $N_e^C = \gamma_2 N_e$.

The hierarchical structure for EPSs of multiple populations can be constructed iteratively as combinations of the two base structures of Fig 2. This construction maintains parameter identifiability. The inference framework for the multiple-populations extension still follows the Inference section: GMRF priors on the base EPS N_e, and the γ_i parameters and the parameter inference is performed with INLA approximations. While we expect many evolutionary scenarios can be expressed in terms of the above hierarchical structures, if the underlying evolutionary process deviates from our model assumption, the model misspecification will lead to bias; formal investigation remains subject to further study.

Although Fig 2 may be interpreted as a realization of a multitype birth-death (MTBD) process [41], our model differs from the approach of MTBD in two important aspects. First, in modeling different genealogical processes, we assume a coalescent process governed by EPS, while MTBD assumes a branching process governed by birth, death, and sampling rates. Second, while both approaches aim to estimate subpopulation rates (EPS in our case, birth/death in MTBD), MTBD additionally targets the locations of the rate changes, whereas our model centers on how the rates of different subpopulations are related to each other over time.

Results

We show the effectiveness of our methodology by applying it to synthetic and real-world data. In the synthetic data section, we offer evidence of its numerical accuracy. The real-data section illustrates the scientific insights that can be obtained by applying our methodology to SARS-CoV-2 data. The code to reproduce the simulation study is available at https://github.com/lorenzocapp/adapop_numexp.

Synthetic data

Fig 3 depicts six pairs of (N_e^A, N_e^B) used to simulate data. The trajectories mimic realistic scenarios typically encountered in applications, such as constant population sizes and exponential growths. The scenarios include several types of dependence between N_e^A and N_e^B, ranging from perfect association (Scenario 6) to no association (Scenario 2). We simulated 100 datasets per each scenario. For a fixed pair of EPSs, we sampled $(\mathbf{s}^A, \mathbf{s}^B)$, $(\mathbf{n}^A, \mathbf{n}^B)$, and $(\mathbf{t}^A, \mathbf{t}^B)$ with $n^A = n^B = 200$. Specifics of (N_e^A, N_e^B) and the data-generating mechanism can be found at Sections B–D in S1 Text.

We refer our hierarchical approach with and without preferential sampling as "adaPop" (Eq 4) and "adaPop+Pref" (Eq 6), respectively, and compare them to the parametric method (Eq 5) referred here as "parPop". We also include a neutral estimator "noPop", which ignores the association between populations and estimates N_e^A and N_e^B independently. We compare how accurately the four methodologies estimate γ, N_e^A and N_e^B. Note that, while parPop and noPop do not approximate γ, the posterior $P(\gamma|\mathbf{s}^A, \mathbf{s}^B, \mathbf{t}^A, \mathbf{t}^B)$ can be empirically approximated by taking samples from $P(N_e^A, N_e^B|\mathbf{s}^A, \mathbf{s}^B, \mathbf{t}^A, \mathbf{t}^B)$ and summarizing $\gamma = N_e^B/N_e^A$. Let f be either γ, N_e^A or N_e^B. We evaluate the performance of the methods using three metrics (listed below) computed on a regular grid of time points $(v_i)_{1:K}$. Here, the grid is defined with $K = 100$ on the interval $[0, 0.6\ T_{MRCA}]$, where T_{MRCA} denotes the time to the most recent common ancestor at the root.

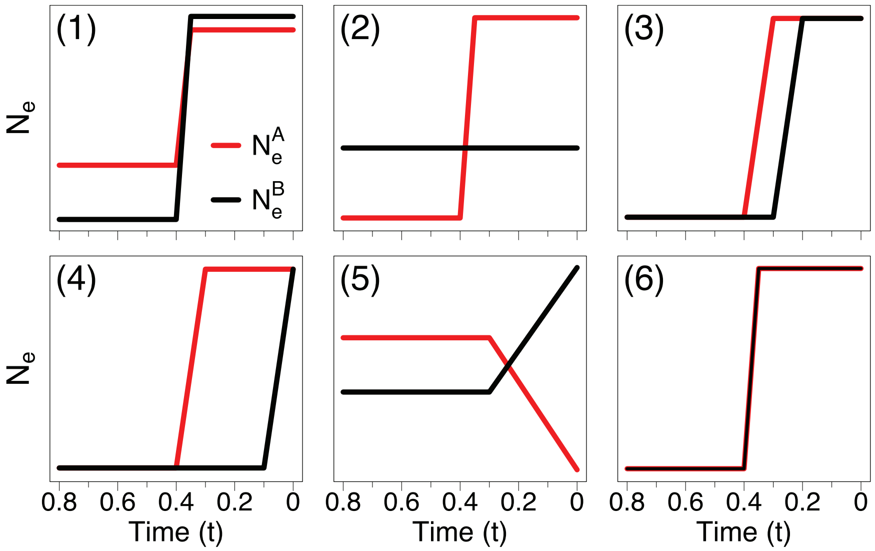

Fig 3. Population trajectories for synthetic data.

https://doi.org/10.1371/journal.pcbi.1010897.g003

- DEV $= \sum_{i=1}^{K} \frac{|\widehat{f}(v_i) - f(v_i)|}{f(v_i)}$, where $\widehat{f}(v_i)$ is the posterior median of f at time v_i. It is a measure of bias.
- RWD $= \sum_{i=1}^{K} \frac{|\hat{f}_{97.5}(v_i) - \hat{f}_{2.5}(v_i)|}{f(v_i)}$, where $\hat{f}_{97.5}(v_i)$ and $\hat{f}_{2.5}(v_i)$ are respectively the 97.5% and 2.5% quantiles of the posterior distribution of $f(v_i)$. It describes the average width of the credible region.
- ENV $= \sum_{i=1}^{K} \mathbb{1}_{\{\hat{f}_{2.5}(v_i) \leq f(v_i) \leq \hat{f}_{97.5}(v_i)\}}$. It is a measure of 95% credible intervals coverage.

Table 1 reports the average value of each statistic pooling together all datasets, all scenarios, and all grid points. Hence, each entry should represent the average performance of a method across the variety of challenging scenarios considered. We also average the performance metrics of N_e^A and N_e^B. A more granular view of the performance of each method by scenario is given in Table A in S1 Text.

Table 1. Summary statistics of posterior inference of γ, N_e^A, and N_e^B. Each entry is computed as the mean of the performance metrics considered across all synthetic datasets: six possible scenarios for (N_e^A, N_e^B), and 100 datasets per a scenario. The metrics for N_e^A and N_e^B have also been averaged. Numbers in parentheses are the standard deviation of each estimate. The numbers in bold indicate the method(s) with the best performance (and within 10% of the best) for each performance metric: the highest for ENV and the lowest for DEV and RWD.

METHOD	$\overline{\text{ENV}}\gamma$	$\overline{\text{DEV}}\gamma$	$\overline{\text{RWD}}\gamma$	$\overline{\text{ENV}}N_e$	$\overline{\text{DEV}}N_e$	$\overline{\text{RWD}}N_e$
adaPop+Pref	**0.89**	**0.34**	**4.41**	**0.93**	**0.3**	**3.15**
	(0.12)	(0.21)	(15.71)	(0.08)	(0.27)	(13.58)
adaPop	**0.93**	**0.36**	6.96	**0.95**	0.39	5.65
	(0.1)	(0.24)	(26.92)	(0.06)	(0.37)	(25.04)
noPop	**0.94**	0.52	90.67	**0.95**	0.34	6.27
	(0.09)	(0.43)	(1144.69)	(0.07)	(0.28)	(24.08)
parPop	0.74	0.55	29.38	0.77	1.03	27.53
	(0.28)	(0.41)	(305.45)	(0.29)	(6.81)	(673.73)

https://doi.org/10.1371/journal.pcbi.1010897.t001

adaPop+Pref stands out as the best performing method, exhibiting the lowest bias (DEV) and the narrowest credible regions (RWD). The coverage (ENV) is slightly worse than noPop; however, noPop achieves the slightly higher coverage with much wider credible regions. adaPop has a very similar performance to adaPop+Pref.

Sampling times were sampled at uniform and not proportionally to N_e^A and N_e^B. This is the case where preferential sampling is less informative. Remarkably, adaPop+Pref still has narrower credible regions than adaPop. The reason is that the time-varying coefficient ζ is able to capture a variety of sampling protocols, including uniform sampling. We interpret this as an empirical evidence of the adaptivity of the model.

adaPop has narrower credible regions than the competing methodologies (noPop, parPop). This is an empirical proof of the "borrowing of information" of a hierarchical model. Notably, this property holds consistently across scenarios. adaPop's higher $\overline{\text{DEV}}N_e$ is mostly attributed to poorer performance in estimating N_e^B in Scenario 2 (see Table A in S1 Text). If we exclude Scenario 2, adaPop is superior to noPop and parPop across all metrics. Similarly, parPop is very competitive in the scenarios where the model is correctly specified (Scenarios 1, 2, and 6). The average performance deteriorates due to poorer performance in the remaining scenarios.

Lastly, an essential feature of adaPop and adaPop+Pref is that they are the most stable methodologies. This can be seen from the standard deviations of the statistics in the parentheses. Table A in S1 Text includes further analyses where the robustness and performance of the models are evaluated.

Real data

Since its introduction, SARS-CoV-2 has undergone rapid evolution resulting in novel variants, some of which possess transmissibility, pathogenicity, or antigenicity advantages over the preexisting resident variants [42, 43]. A variant of recent interest is the delta variant (Pango lineage B.1.617.2 and AY lineages [44]). We compare this variant to other SARS-CoV-2 variants in two countries, South Korea and Italy.

We analyzed high-coverage complete sequences publicly available in GISAID [45] collected from South Korea and Italy during 2021-03-01 to 2021-09-30. For each country, we subsampled two sets of 150 sequences: one with the delta variant and the other without the delta variant. The details of the sequences used for our analysis can be found at Table B and Fig A in S1 Text. We then estimated the maximum credibility clade (MCC) trees—the tree in the posterior sample with the maximum sum of the posterior clade probabilities—of samples from each variant group of each country independently with BEAST2 [3]; the further analysis pipeline can be found at Section E in S1 Text. In our study, we set the population of the delta variant sequences as population A and of the non-delta variant as population B. Here, we discuss the results using the parPop and adaPop+Pref models. The additional results with other methods and further details of the sequence analysis pipeline, together with the inferred MCC trees, appear in Figs B–F in S1 Text.

The orange-shaded heatmaps in Fig 4 depict the number of samples collected over time of the two populations in South Korea and Italy (as represented in our sub-sampled data sets). The observed pattern is consistent across the two countries: delta viral samples were predominantly collected in the summer of 2021, while non-delta samples were collected in the spring and early summer of 2021. This is consistent with the general observation of rapid spread of the delta variant that has progressively replaced the preexisting non-delta variants (such as alpha) since its introduction [46, 47].

Fig 4B, 4C, 4F and 4G depict the estimated posterior distribution of EPS of the two viral populations in the two countries obtained with adaPref+Pop (noPop estimates are qualitatively

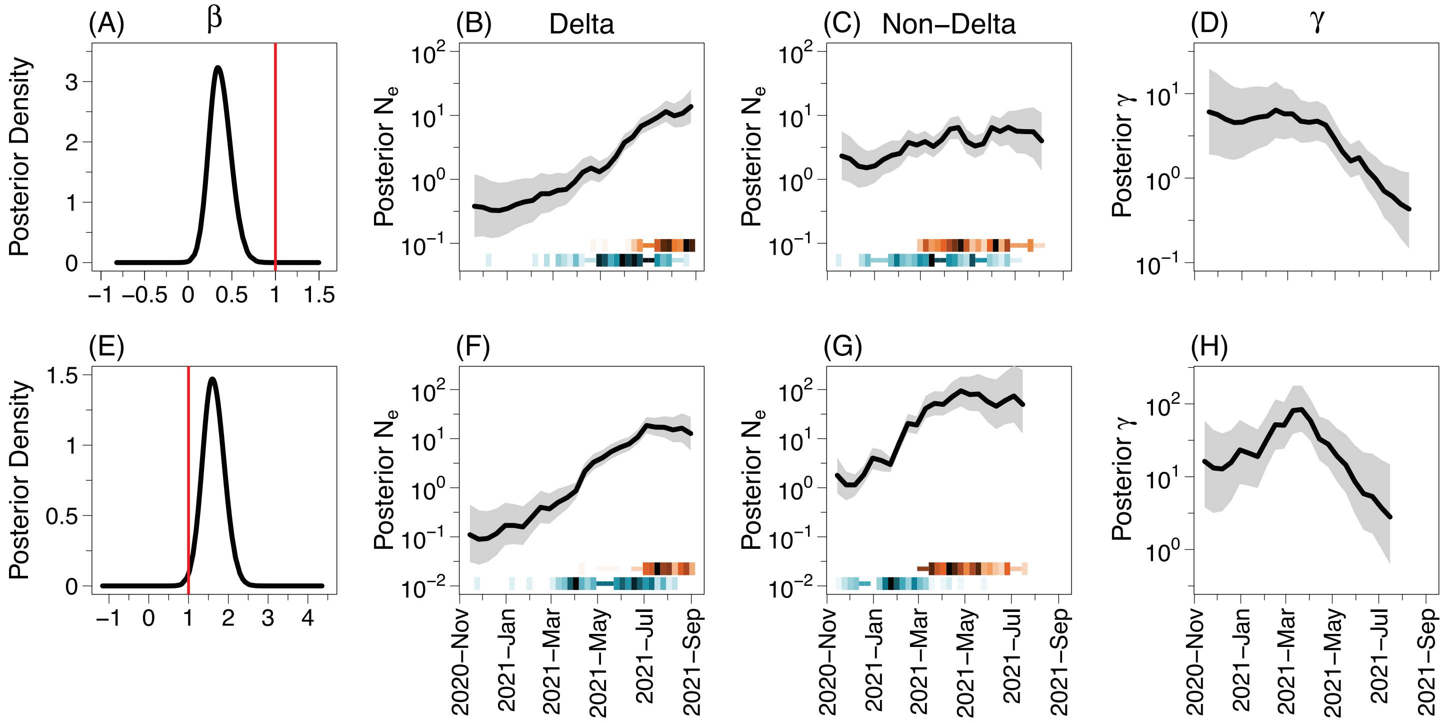

Fig 4. Posterior inference of SARS-CoV-2 population dynamics in South Korea and Italy. Panels A–D contain results of South Korea, and the panels E–H show results of Italy with $\mathbf{g}^A$=delta and $\mathbf{g}^B$=non-delta. The first column shows the β parameter posterior density from the parPop method. The red line indicates the value of β under the hypothesis that both variants share the same EPS trajectory in the parPop model. The other columns present the results with adaPop+Pref: the second and the third columns display posterior estimates of EPS of the delta and non-delta variant, respectively, and the last column shows posterior estimates of γ. The solid line indicates the posterior medians with its surrounding shaded areas representing 95% BCIs. The orange and blue heatmaps describe the sampling and coalescent event intensity, respectively: the darker the color, the more number of events occurs in a time interval. The y-axis of plots in the columns 2–4 is plotted on a log scale. The results using other methods can be found in the Figs D–F in S1 Text.

https://doi.org/10.1371/journal.pcbi.1010897.g004

similar, see Fig D in S1 Text). In both countries, the delta population has experienced prolonged growth since its inception and halted its growth in the last month. On the other hand, the non-delta population EPS grew until approximately the end of May (the growth looks more pronounced in Italy), then it roughly plateaued.

We next examine the quantification of the dependence between the two populations: β for parPop and γ for adaPop+Pref. In Fig 4A, the posterior density of β estimated under the parPop model has mean 0.36 with 95% credible interval (0.19, 0.51), well below 1 (red line), suggesting EPS growth is more pronounced among sequences having the delta variant in South Korea. On the other hand, the posterior density of β has mean 1.62 with credible interval (1.25, 1.95) (Fig 4E) indicating the EPS of the non-delta variant grows faster than the delta variant in Italy under the parPop model. The result suggests that the growth of the non-delta population dominated that of the delta population which contradicts the general consensus [46, 47].

Under the adaPop+Pref model, the monotonic decrease in the growth of the non-delta variant EPS compared to that of the delta variant EPS in South Korea is apparent in γ-trajectory (Fig 4D); this is consistent with parPop β. However, the γ-trajectory of Italy (Fig 4H) shows that, compared to the growth of the delta variant EPS, the non-delta variant underwent an initial phase of faster growth and then transitioned to slower growth around mid-March of 2021. The inability of the parPop model (Fig 4E) to capture the two-phase population dynamics in Italy, that are evident in the adaPop+Pref model (Fig 4H), suggests that more flexible approaches proposed by our work are needed for accommodating the broad range of population dynamics scenarios encountered in real applications. We interpret the discrepancy between β and γ in this case as the evidence that parPop is not correctly specified.

Discussion

We have developed a coalescent-based Bayesian methodology for inferring dependent population size trajectories and quantifying such dependency. We make minimal assumptions on the functional form of population size trajectories and allow the dependence between the two populations to vary over time. We also present a sampling-aware model for leveraging additional information contained in sampling times for reduced bias and improved inference accuracy. Although the proposed models have an increased number of parameters, we prove that the models are identifiable. We have shown that our adaPop+Pref outperforms other methods in synthetic data with known ground truth and that our adaptive method can detect changes in population size dynamics that are otherwise undetected with other models. We make this point more precise in our SARS-CoV-2 analyses.

We have decided to infer parameters via a numerically approximated method that relies on Laplace approximations (INLA) for computational speed. However, our proposed models can, in principle, be implemented in any of the MCMC standard approaches. Implementing our models in an MCMC approach would allow to infer population size trajectories from molecular sequences and sequencing time information directly and it is a subject of future development. This extension would account for uncertainty in the genealogy, an important component that is missing in the analyses presented.

We see a growing number of applications of the coalescent requiring the modeling of complex demographic histories. In the introduction, we mentioned a few, such as viral epidemiology and cancer evolution. There is extensive literature on models incorporating more and more realistic features, for example, a detailed description of migration histories. We note however that our proposed model explicitly models the dependency of population trajectories, providing a more interpretable dependency than in the structured coalescent. The quest for realism and scientifically meaningful parameters comes at the cost of computational tractability and leads, sometimes, to issues of model identifiability. Our work is somewhat motivated by these problems. Our method is equipped with high performance and accuracy, due to its scalability, interpretability, and parameter identifiability; such properties are lacking in many complex models in biology and epidemiology. We see our proposal as a “hybrid approach” that allows scientists to quantify the relative advantage of one population over another while still retaining a fairly parsimonious model. Such an approach will be invaluable across many biomedical disciplines for studying complex time-varying dependent evolutionary dynamics of populations.

Supporting information

S1 Text. Section A. Proofs of identifiability. Section B. Simulation details. Section C. Grid construction. Section D. Synthetic data: additional results. Section E. SARS-CoV-2 molecular data analysis. Table A. Summary statistics of posterior inference of γ, N_e^A, and N_e^B. Each entry is computed as the mean of the performance metric for a given scenario (100 datasets per scenario). The metrics for N_e^A and N_e^B have been also averaged. The numbers in bold indicate the method(s) with the best performance (and within 10% of the best) for each performance metric: the highest for ENV, the lowest for DEV and RWD. **Table B. GISAID EPI_SET IDs and their corresponding DOIs for sequences used in the real data analysis.** Each EPI_SET dataset contains 150 sequences. Note that EPI_ISL_402124 (hCoV-19/Wuhan/WIV04/2019, the official reference sequence employed by GISAID) is automatically included

in the DOI EPI_SET web viewer generated by GISAID in addition to the 150 sequences per each dataset below. **Fig A. Collection date distributions of available high-coverage complete sequences in GISAID.** (A) South Korea. (B) Italy. Purple and green colors indicate sequences with and without delta variants, respectively. **Fig B. MCC trees of delta and non-delta variants from South Korea. Fig C. MCC trees of delta and non-delta variants from Italy. Fig D. Posterior EPS trajectories using the noPop method.** (A) South Korea, delta EPS. (B) South Korea, non-delta EPS. (C) Italy, delta EPS. (D) Italy, non-delta EPS. The figure format follows Fig 4 of the main text. **Fig E. Posterior densities of parameters and posterior EPS trajectories using the parPop method.** (A) South Korea, log α. (B) South Korea, β. (C) South Korea, delta EPS. (D) South Korea, non-delta EPS. (E) Italy, log α. (F) Italy, β. (G) Italy, delta EPS. (H) Italy, non-delta EPS. The figure format follows Fig 4 of the main text. **Fig F. Posterior EPS trajectories and posterior estimates of γ using the adaPop method.** (A) South Korea, delta EPS. (B) South Korea, non-delta EPS. (C) South Korea, γ, (D) Italy, β. (E) Italy, delta EPS. (F) Italy, γ. The figure format follows Fig 4 of the main text.
(PDF)

Author Contributions

Conceptualization: Lorenzo Cappello, Jaehee Kim.

Data curation: Lorenzo Cappello, Jaehee Kim.

Formal analysis: Lorenzo Cappello, Jaehee Kim.

Funding acquisition: Julia A. Palacios.

Investigation: Lorenzo Cappello, Jaehee Kim.

Methodology: Lorenzo Cappello, Jaehee Kim.

Project administration: Lorenzo Cappello, Jaehee Kim.

Resources: Lorenzo Cappello, Jaehee Kim.

Software: Lorenzo Cappello.

Supervision: Julia A. Palacios.

Validation: Lorenzo Cappello, Jaehee Kim.

Visualization: Lorenzo Cappello, Jaehee Kim.

Writing – original draft: Lorenzo Cappello, Jaehee Kim, Julia A. Palacios.

Writing – review & editing: Lorenzo Cappello, Jaehee Kim, Julia A. Palacios.

References

1. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution. 2005; 22(5):1185–1192. https://doi.org/10.1093/molbev/msi103 PMID: 15703244
2. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of infectious disease epidemics. Genetics. 2009; 183(4):1421–1430. https://doi.org/10.1534/genetics.109.106021 PMID: 19797047
3. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology. 2019; 15(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650 PMID: 30958812
4. Stadler T, Pybus OG, Stumpf MPH. Phylodynamics for cell biologists. Science. 2021; 371(6526): eaah6266. https://doi.org/10.1126/science.aah6266 PMID: 33446527

5. Cappello L, Kim J, Liu S, Palacios JA. Statistical challenges in tracking the evolution of SARS-CoV-2. Statistical Science. 2022; 37(2):162–182. https://doi.org/10.1214/22-sts853 PMID: 36034090

6. Featherstone LA, Zhang JM, Vaughan TG, Duchene S. Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications. Virus Evolution. 2022; 8(1):veac045. https://doi.org/10.1093/ve/veac045 PMID: 35775026

7. Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. Molecular Biology and Evolution. 2016; 33(8):2102–2116. https://doi.org/10.1093/molbev/msw064 PMID: 27189573

8. Müller NF, Rasmussen DA, Stadler T. The structured coalescent and its approximations. Molecular Biology and Evolution. 2017; 34(11):2970–2981. https://doi.org/10.1093/molbev/msx186 PMID: 28666382

9. Caswell-Jin JL, Lorenz C, Curtis C. Molecular heterogeneity and evolution in breast cancer. Annual Review of Cancer Biology. 2021; 5:79–94. https://doi.org/10.1146/annurev-cancerbio-060220-014137

10. Kingman JFC. The coalescent. Stochastic Processes and Their Applications. 1982; 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4

11. Wakeley J, Sargsyan O. Extensions of the coalescent effective population size. Genetics. 2009; 181 (1):341–345. https://doi.org/10.1534/genetics.108.092460 PMID: 19001293

12. Ho SYW, Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular Ecology Resources. 2011; 11(3):423–434. https://doi.org/10.1111/j.1755-0998.2011.02988.x PMID: 21481200

13. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution. 2008; 25(7):1459–1471. https://doi.org/10.1093/molbev/msn090 PMID: 18408232

14. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular Biology and Evolution. 2013; 30(3):713–724. https://doi.org/10.1093/molbev/mss265 PMID: 23180580

15. Volz EM, Didelot X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. Systematic Biology. 2018; 67 (4):719–728. https://doi.org/10.1093/sysbio/syy007 PMID: 29432602

16. Faulkner JR, Magee AF, Shapiro B, Minin VN. Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. Biometrics. 2020; 76(3):677–690. PMID: 32277713

17. Palacios JA, Minin VN. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. Biometrics. 2013; 69(1):8–18. https://doi.org/10.1111/biom.12003 PMID: 23409705

18. Adams RP, Murray I, MacKay DJ. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009;9–16.

19. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018; 4(1):vey016. https://doi.org/10.1093/ve/vey016 PMID: 29942656

20. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software. 2017; 76(1):1–32. https://doi.org/10.18637/jss.v076.i01 PMID: 36568334

21. Lan S, Palacios JA, Karcher M, Minin VN, Shahbaba B. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. Bioinformatics. 2015; 31(20):3282–3289. https://doi.org/10.1093/bioinformatics/btv378 PMID: 26093147

22. Shahbaba B, Lan S, Johnson WO, Neal RM. Split Hamiltonian Monte Carlo. Statistics and Computing. 2014; 24(3):339–349. https://doi.org/10.1007/s11222-012-9373-1

23. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B. 2009; 71 (2):319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

24. Palacios JA, Minin VN. Integrated Nested Laplace Approximation for Bayesian Nonparametric Phylodynamics. In: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. 2012;726–735.

25. Volz EM, Frost SDW. Sampling through time and phylodynamic inference with coalescent and birth & death models. Journal of the Royal Society Interface. 2014; 11(101):20140945. https://doi.org/10.1098/rsif.2014.0945 PMID: 25401173

26. Diggle PJ, Menezes R, Su T. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C. 2010; 59(2):191–232.

27. Karcher MD, Palacios JA, Bedford T, Suchard MA, Minin VN. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLOS Computational Biology. 2016; 12(3): e1004789. https://doi.org/10.1371/journal.pcbi.1004789 PMID: 26938243

28. Karcher MD, Carvalho LM, Suchard MA, Dudas G, Minin VN. Estimating effective population size changes from preferentially sampled genetic sequences. PLOS Computational Biology. 2020; 16 (10):1–22. https://doi.org/10.1371/journal.pcbi.1007774 PMID: 33044955

29. Parag KV, du Plessis L, Pybus OG. Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. Molecular Biology and Evolution. 2020; 37(8):2414–2429. https://doi.org/10.1093/molbev/msaa016 PMID: 32003829

30. Cappello L, Palacios JA. Adaptive preferential sampling in phylodynamics with an application to SARS-CoV-2. Journal of Computational and Graphical Statistics. 2021; 0:1–29. https://doi.org/10.1080/10618600.2021.1987256 PMID: 36035966

31. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. New York, NY: Chapman and Hall/CRC; 1995.

32. Faulkner JR, Minin VN. Locally adaptive smoothing with Markov random fields and shrinkage priors. Bayesian Analysis. 2018; 13(1):225. https://doi.org/10.1214/17-BA1050 PMID: 29755638

33. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell. 2021; 184(1):64–75. https://doi.org/10.1016/j.cell.2020.11.020 PMID: 33275900

34. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021; 372 (6538). https://doi.org/10.1126/science.abg3055 PMID: 33658326

35. Rothenberg TJ. Identification in parametric models. Econometrica. 1971; 39:577–591. https://doi.org/10.2307/1913267

36. Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer; 2006.

37. Watanabe S. Algebraic Geometry and Statistical Learning Theory. Cambridge, UK: Cambridge University Press; 2009.

38. Little MP, Heidenreich WF, Li G. Parameter identifiability and redundancy: theoretical considerations. PLOS One. 2010; 5(1):e8915. https://doi.org/10.1371/journal.pone.0008915 PMID: 20111720

39. Parag KV, Pybus OG. Robust design for coalescent model inference. Systematic Biology. 2019; 68 (5):730–743. https://doi.org/10.1093/sysbio/syz008 PMID: 30726979

40. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000; 155(3):1429–1437. https://doi.org/10.1093/genetics/155.3.1429 PMID: 10880500

41. Barido-Sottani J, Vaughan TG, Stadler T. A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. Systematic Biology. 2020; 69(5):973–986. https://doi.org/10.1093/sysbio/syaa016 PMID: 32105322

42. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology. 2021; 19(7):409–424. https://doi.org/10.1038/s41579-021-00573-0 PMID: 34075212

43. Cevik M, Grubaugh ND, Iwasaki A, Openshaw P. COVID-19 vaccines: Keeping pace with SARS-CoV-2 variants. Cell. 2021; 184(20):5077–5081. https://doi.org/10.1016/j.cell.2021.09.010 PMID: 34534444

44. 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. https://doi.org/10.1038/s41564-020-0770-5 PMID: 32669681

45. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017; 22(13):30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMID: 28382917

46. Mlcochova P, Kemp SA, Dhar MS, et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. Nature. 2021; 599(7883):114–119. https://doi.org/10.1038/s41586-021-03944-y PMID: 34488225

47. del Rio C, Malani PN, Omer SB. Confronting the Delta Variant of SARS-CoV-2, Summer 2021. JAMA. 2021; 326(11):1001–1002. https://doi.org/10.1001/jama.2021.14811 PMID: 34406361