Inference with selection, varying population size, and evolving population structure: application of ABC to a forward–backward coalescent process with interactions

Clotilde Lepers¹ · Sylvain Billiard ² · Matthieu Porte³ · Sylvie Méléard⁴ · Viet Chi Tran ⁵

Received: 24 October 2019 / Accepted: 15 October 2020 / Published online: 30 October 2020
© The Author(s), under exclusive licence to The Genetics Society 2020

Abstract

Genetic data are often used to infer demographic history and changes or detect genes under selection. Inferential methods are commonly based on models making various strong assumptions: demography and population structures are supposed a priori known, the evolution of the genetic composition of a population does not affect demography nor population structure, and there is no selection nor interaction between and within genetic strains. In this paper, we present a stochastic birth-death model with competitive interactions and asexual reproduction. We develop an inferential procedure for ecological, demographic, and genetic parameters. We first show how genetic diversity and genealogies are related to birth and death rates, and to how individuals compete within and between strains. This leads us to propose an original model of phylogenies, with trait structure and interactions, that allows multiple merging. Second, we develop an Approximate Bayesian Computation framework to use our model for analyzing genetic data. We apply our procedure to simulated data from a toy model, and to real data by analyzing the genetic diversity of microsatellites on Y-chromosomes sampled from Central Asia human populations in order to test whether different social organizations show significantly different fertilities.

Associate editor: Giorgio Bertorelle

Supplementary information The online version of this article (<https://doi.org/10.1038/s41437-020-00381-x>) contains supplementary material, which is available to authorized users.

- ✉ Clotilde Lepers
clotilde.lepers@inserm.fr
- ✉ Sylvain Billiard
sylvain.billiard@univ-lille.fr
- ✉ Matthieu Porte
matthieu.porte@ign.fr
- ✉ Sylvie Méléard
sylvie.meleard@polytechnique.edu
- ✉ Viet Chi Tran
chi.tran@univ-eiffel.fr

- ¹ Université de Paris, INSERM, IAME, F-75006 Paris, France
- ² Univ. Lille, CNRS, UMR 819 8 -Evo-Eco-Paleo, F-59000 Lille, France
- ³ IGN, Institut National de l'Information Géographique et Forestière, F-94165, Saint-Mandé, France
- ⁴ CMAP, CNRS, Ecole Polytechnique, Institut polytechnique de Paris, route de Saclay, 91128 Palaiseau Cedex, France
- ⁵ LAMA, Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, F-77454 Marne-la-Vallée, France

Introduction

Demographic, spatial or genetic structures affect genetic diversity because they determine genetic flows between lineages, relationships between individuals, and coalescent rates (Charlesworth et al. 2003). In turn, genetic polymorphism within and between taxa is commonly used for estimating population structures (Goldstein and Chikhi 2002, Müller et al. 2017) or demographic changes (Beichman et al. 2018), to infer population history, migration patterns, or to search for genes under selection (Stephan 2016). These methods are mostly based either on the site frequency spectrum, the identity per state or descent, or on summary statistics in an Approximate Bayesian Computation (ABC) framework (Beaumont et al. 2002).

Statistical testing and model selection are generally performed under simplifying assumptions which allow computations of quantities such as the likelihood of a model, in particular under neutrality. For instance, under the Wright-Fisher model, the population size is supposed deterministic: it is known at any given time and independent of the composition of the population, i.e., it is supposed that the mechanisms underlying the variations of the population size are extrinsic and without noise. Individuals thus

compete for space but the carrying capacity of the environment does not change because of the evolution of the population itself, or because of extrinsic or intrinsic stochasticity. In birth-death models, population size can vary but populations can grow indefinitely because individuals do not interact. In addition, the Wright–Fisher and birth-death models are most often supposed neutral when used for demographic inference, i.e., the reproduction and survival rates do not depend on the genetic lineage (but see a recent birth-death model without interactions where rates can depend on mutations Rasmussen and Stadler 2019).

Yet, the assumptions of neutrality, extrinsic control of population size or non-interacting individuals are certainly often violated. For instance, genealogies of the seasonal influenza virus show important departure from neutrality which might suggest that selection and interaction between lineages are important enough to significantly affect evolution and the shapes of the phylogenetic trees (Bedford et al. 2011, Strelkowa and Lässig 2012). Reproduction rates and carrying capacities have also been shown to depend on strains in the domesticated yeasts (Spor et al. 2009), and the ecological literature contains many cases where competitive interactions vary among strains or species (Gallieni 2017). Finally, not explicitly including competition in spatially structured population leads to biological inconsistencies in population genetics models (Felsenstein 1975). Developing models and inference methods which relax such hypotheses is thus a contemporaneous challenge, in order to improve our knowledge of the history and ecological features of species and populations. As emphasized by Frost et al. (2015), this challenge is particularly important for the analysis of phylodynamics in clonal species such as viruses.

Some of these assumptions have been already relaxed. For instance, Rasmussen and Stadler (2019) developed a model where reproductive and death rates can differ between lineages which can emerge because of spontaneous mutations. They applied their method on Ebola and influenza viruses in order to have estimates of fitness effects of mutations from phylodynamics. Indeed, variation of death and birth rates between lineages can affect viruses phylogenies, which can be detected and used to infer the effect of mutations. However, they supposed no interaction between lineages, discarding a possible effect of competition between viruses strains.

In this paper, we present a model and an inference method which allow the relaxation of several of these assumptions. First, in section “Genetic diversity in an eco-evolutionary dynamics with three timescales: The substitution Fleming–Viot process (SFVP)”, we recall the stochastic process describing the eco-evolution of a structured population with ecological feedbacks (introduced in Billiard et al. 2015). This model takes into account: (i) A trait

structure that can affect birth, death and competitive rates. The traits, which evolve because of mutations and selection, are seen as proxies for the species, taxa or strains; (ii) Explicit competitive interactions between and within lineages; (iii) Varying population sizes depending on the genetic composition of the population, i.e., the carrying capacity depends on the ecological properties of existing strains (their birth, death, and competitive rates). The model assumes that reproduction is asexual, that mutations affecting fitness are rare, and that neutral mutation follows an intermediate timescale between reproduction and death rates (the ecological timescale) and the rate at which mutations affecting fitness appear (the evolutionary timescale). Second, in section “Genealogies in a forward–backward coalescent with competitive interactions”, a new forward–backward coalescent process is proposed to describe the phylogenies in such a population. The forward step accounts for interactions, demography and evolution of trait structures, defining the skeleton on which the phylogenies of sampled individuals can be reconstructed in the backward step. Phylogenies of structured populations have been previously modeled in nested coalescent models (e.g. Benitez et al. 2018, 2020, Duchamps 2018, Verdu et al. 2009) but, in our case, interactions within and between lineages, ecological feedbacks between selection and population size, and multiple coalescence mergers, are taken into account. Contrarily to Λ -coalescent models proposed in the literature (Donnelly and Kurtz 1999, Pitman 1999, Sagitov 1999), multiple merging here are not due to sweepstakes reproductive successes but they appear as a consequence of natural selection via mutation-competition and timescales. Third, in section “ABC inference in an eco-evolutionary framework”, we develop an ABC framework in order to estimate the parameters of the model from genetic diversity data. We show how ecological parameters such as individual birth and death rates, and competitive abilities can be estimated. Finally, we apply our inferential procedure on simulated data from an eco-evolutionary toy model, and on genetic data from Y-chromosomes sampled in Central Asia human populations (Chaix et al. 2007, Heyer et al. 2015) in order to test whether different social organizations can be associated with differences in fertility.

The forward–backward coalescent model

In the current work, we extend the population model developed in Billiard et al. (2015) (following Champagnat 2006, Champagnat and Méléard 2007, Metz et al. 1996) to include phylogenies and develop a statistical ABC procedure that we apply on simulated and real datasets. The eco-evolution of a structured population with ecological feedbacks is described by a stochastic process. The population is

structured by traits, considered as proxies for species, taxa or strains. These traits can affect birth, death and competitive rates, and new traits are generated by mutations. Explicit competitive interactions are modeled between individuals of the population with intensities depending on the traits, inducing varying population sizes depending on the genetic composition of the population. Also, a marker structure is added. Markers are assumed neutral in the sense that they have no impact on fecundity, survival or competition. They are introduced in the model to measure the neutral diversity and allow the reconstruction of the phylogenies. The model assumes asexual reproduction and complete linkage between traits and markers, and that the population evolves following three timescales. First, the ecological timescale: birth and death rates occur at a fast rate. Second, marker mutations arise slightly slower than the ecological timescale. Finally, mutations on the trait under selection occur at the slowest timescale. This reflects for instance that a large proportion of a genome is not composed of traits under selection. This happens for example in the influenza virus which shows a large diversity within seasons despite a very rapid evolution and adaptation (Neher and Bedford 2015).

Before precisely describing the application of the model to infer demographic and genetic parameters within an ABC framework, we summarize hereafter the main features and outcomes of the model.

Genetic diversity in an eco-evolutionary dynamics with three timescales: the substitution Fleming–Viot process (SFVP)

We assume a population of clonal individuals characterized, on the one hand, by a trait $x \in \mathcal{X} \subset \mathbb{R}^d$, which affects the demographic processes such as birth, death, and competitive interactions between individuals and, on the other hand, by a vector of genetic markers $u \in \mathcal{U} \subset \mathbb{R}^q$, supposed neutral (i.e., u does not affect the demographic process). Individuals with trait x give birth at rate $b(x)$, and $d(x)$ is their intrinsic death rate. The competitive interactions between individuals with traits x and y add an effect $C(x, y)$ on the individual death rate. When the population is large, the evolution of the population can be decomposed into the succession of invasions of favorable mutations on the trait x , because ecological processes are very fast, and the population jumps from one state to another. The neutral marker also evolves between each adaptive jump, at a faster timescale that is compensated by mutations of small effect. Since the ecological parameters change after each adaptive jump on trait x (the birth rate, death rate and the population size change), the evolution of the neutral marker also changes. Hence, even if the marker is neutral, its own evolution

depends on the state of the population at a given time, especially on the competitive interactions $C(x, y)$ between individuals with traits x and y . Overall, the joint eco-evolutionary dynamics of the neutral marker and the selected traits can be approximated by the so-called Substitution Fleming–Viot Process (SFVP, (Billiard et al. 2015), see Appendix A in supplementary materials for details).

Distribution of the trait x between two adaptive jumps

At the ecological timescale, when the population is large, p strains with traits $x_1, \dots, x_p$ can coexist. Between two adaptive jumps, the trait distribution in the population remains almost constant. Indeed, the size of subpopulations can vary but are expected to stay close to their equilibria $\hat{n}(x_1; x_1, \dots, x_p), \dots, \hat{n}(x_p; x_1, \dots, x_p)$, given by the following competitive Lotka–Volterra system of ordinary differential equations (ODE) that approximates the evolution in the ecological timescale:

$$\frac{dn_t(x_j)}{dt} = (b(x_j) - d(x_j) - \sum_{\ell=1}^p C(x_j, x_\ell) n_t(x_\ell)) n_t(x_j), \quad j \in \{1, \dots, p\}, \quad (2.1)$$

where $n_t(x)$ can be seen as the density of individuals of strain with trait x . The equilibrium $\hat{n}(x_i; x_1, \dots, x_p)$ of the strain with trait x_i depends on the whole trait structure of the population which is in turn defined entirely by the set of traits present in the population (the arguments of $\hat{n}$ given after the semicolon).

Change of the distribution of the trait x during an adaptive jump

In the timescale of trait mutations occurring in a population composed of p strains with traits $x_1, \dots, x_p$ and respective sizes $\hat{n}(x_1; x_1, \dots, x_p), \dots, \hat{n}(x_p; x_1, \dots, x_p)$, when a mutation on trait x_i occurs at time t , a new strain is introduced with trait $x_i + h$ where h is drawn in a distribution $m(x_i, h)dh$ (mutations on trait x are not necessarily small, i.e., selection can be strong). Whether the mutant strain invades or not the population depends on its invasion fitness defined by

$$f(y; x_1, \dots, x_p) = b(y) - d(y) - \sum_{j=1}^p \hat{n}(x_j; x_1, \dots, x_p) C(y, x_j) \quad (2.2)$$

(Champagnat, 2006, Champagnat et al. 2006, Metz et al. 1996). The mutant strain invades with probability $\frac{[f(x_i+h; x_1, \dots, x_p)]_+}{b(x_i+h)}$, in which case the population jumps to a new state given by the solution of the Lotka–Volterra ODE

system (Eq. (2.1)) updated with the introduction of the mutant strain $(\hat{n}(x_1; x_1, \dots, x_p, x_i + h), \dots, \hat{n}(x_i + h; x_1, \dots, x_p, x_i + h))$. In the new equilibrium, some former traits $x_1, \dots, x_p$ may be lost. The evolution of the trait can thus be described by a Polymorphic Evolution Sequence (PES), i.e., the succession of the adaptive jumps of the population from one state to another (Champagnat and Méléard 2011). For a visual abstract of the PES, see Fig. A.1 in supplementary materials.

Evolution of the neutral marker

When the mutant strain with trait $x = x_i + h$ invades the population, say at time 0, an adaptive jump occurs. Let us denote by u the marker of the first mutant individual (x, u) . Initially, the distribution of the neutral marker within strain i and trait x , is thus composed of a single individual with marker u . The evolution of the marker distribution within this strain is given by $F_t^u(x, dv)$, the distribution at time t of the marker values within the strain with trait x given the initial value u . This distribution changes with time depending on the supposed mutation kernel on the marker, on the birth and death rates of individuals with trait x , and on the competitive interactions $C(x, y)$ with all the other individuals of any trait value $y \in \{x_1, \dots, x_p, x_i + h\}$. Between two adaptive jumps, assuming small marker mutations but not necessarily small trait mutations, how the distribution $F_t^u(x, dv)$ evolves with time is given by the following stochastic differential equation (see (Billiard et al. 2015)) (derivation details and a more general form are given in Appendix A in supplementary materials)

$$\int_{\mathcal{U}} \phi(v) F_t^u(x, dv) = \phi(u) + b(x) \int_0^t \left(\int_{\mathcal{U}} \Delta \phi(v) F_s^u(x, dv) \right) ds + M_t^x(\phi). \quad (2.3)$$

The left side of the equation can be seen as the expectation of the distribution of the marker value at time t , where ϕ is a test function (supposed twice differentiable on $\mathcal{U}$). Different choices of functions ϕ will provide descriptors of the distribution F_t^u (for example $\phi(v) = v$ gives the mean of the distribution). The right side of the equation tells what is the expected form of the distribution. The first term on the right side gives the initial conditions: the first mutant with trait x has a marker value u , hence the initial condition for the distribution is $\phi(u)$. The second term on the right side integrates the changes of the distribution which are only due to mutations on the marker between time 0 (the invasion time of x) and t . Since mutation only occurs at birth, the rate at which F changes with mutation is proportional to the birth rate $b(x)$. Within the integral, $\Delta \phi(v)$ is the Laplacian of the function ϕ which gives the rate of change of F in all the dimensions of the marker values (which depends on the

assumptions made on the mutation kernel and can be generalized, see Appendix A in supplementary materials). The last term $M_t^x(\phi)$ on the right side gives the changes of F which are due to the ecological processes, i.e., the fluctuations due to the birth and death of the individuals with trait x . $M_t^x(\phi)$ is a martingale i.e., a square integrable random variable with mean 0 and variance

$$\text{Var}(M_t^x(\phi)) = \frac{2b(x)}{\hat{n}(x; x_1, \dots, x_p, x_i + h)} \times \int_0^t \mathbb{E} \left[\left(\int_{\mathcal{U}} \phi^2(v) F_s^u(x, dv) - \left(\int_{\mathcal{U}} \phi(v) F_s^u(x, dv) \right)^2 \right) ds \right]. \quad (2.4)$$

The fraction in the right hand side (r.h.s.) of Eq. (2.4) corresponds to the demographic variance $2b(x)$ divided by the effective population size

$$N_e(x) = \hat{n}(x; x_1, \dots, x_p, x_i + h). \quad (2.5)$$

The population effective size, which partially governs the evolution of the diversity at the neutral marker, depends on the trait value x , but also on the whole trait distribution $x_1, \dots, x_p, x_i + h$. In particular, it means that the variance in the neutral diversity within the strain with trait x depends on the competitive interactions of the latter with all the other strains.

Genealogies in a forward-backward coalescent with competitive interactions

Genealogies are piecewise-defined and constructed by dividing time between intervals separating adaptive jumps of the PES, following a forward-backward coalescent process. Since the evolution of trait x depends on the current distribution of the traits in the population, the PES tree is constructed forward in time where the successive adaptive jump times are denoted by $(T_k)_{k \in \{1, \dots, J\}}$, with $T_0 = 0$ and J is the number of jumps that occurred before time t . During the PES, a subpopulation with trait x_i has its own coalescent rate on the markers which depends on its reproductive rate $b(x_i)$ and on the distribution of the traits in the whole population (Eq. (2.5)). Genealogies are thus expected to be different among the different strains and between different adaptive jumps of the PES. Between adaptive jumps, since under our assumptions trait x distribution and population size are supposed fixed, within-strains genealogies can be constructed backward in time. Given the PES during the time interval $[T_k, T_{k+1})$ ($k \in \{0, \dots, J-1\}$) and the trait distribution $\{x_1, \dots, x_p\}$, the genealogy of the individuals within the strain with trait x_i is obtained by simulating a Kingman coalescent with coalescence rate $\frac{2b(x_i)}{\hat{n}(x_i; x_1, \dots, x_p)}$ (Eq. (2.4)). When an adaptive jump

occurs at time T_k , all lineages in the subpopulation of strain x_i instantaneously coalesce because a single mutant is always at the origin of a new strain during the PES. Note that coalescence is instantaneous under the assumptions underlying the PES, i.e., at the timescale governing the evolution of the trait, the transition to fixation of the mutant trait is negligible. The allelic state at the marker is determined given the previously constructed genealogy, depending on the mutational model considered.

A more formal definition of the coalescent and associated proofs are given in Appendix A.3 (see supp. mat.). A simulation algorithm for the construction of genealogies under our model is given in Appendix A.4 (see supp. mat.).

ABC inference in an eco-evolutionary framework

We showed in the previous sections that the genetic structure of a sample of n individuals can be related to the parameters of our eco-evolutionary model. We now aim at using this framework to infer genealogies, ecological and genetic parameters from genetic and/or phenotypic data sampled in a population at time t . In other words, given a dataset containing the genotype at the marker u and the genotype or phenotype at the trait x for the n sampled individuals, we want to infer the parameters of the model: birth, death and competitive interaction rates, mutation rates, etc. Since we have only a partial information on the population (n individuals are sampled and possible extinct lineages are unobserved), the likelihood of a model given the data have no tractable form. Indeed, given a possible genealogy of the n individuals, an infinite number of continuous genealogical trees could be obtained from the model. The likelihood of each tree depends on the number and the traits of the different subpopulations (or strains) during the history of the population, including the unobserved and extinct ones. Because summing over all possible unobserved data (number of unobserved and extinct lineages with their traits and adaptive jump times) is not feasible in practice, we have to make inference without likelihood computations.

An alternative to likelihood-based inference methods is given by the Approximate Bayesian Computation (ABC) (Beaumont et al. 2009, 2002), which relies on repeated simulations of the forward-backward coalescent trees (section “Genealogies in a forward-backward coalescent with competitive interactions”). In the following, we briefly give a general presentation of the application of the ABC method to our model. We then apply the method to simulations of a toy model (the Dieckmann-Doebeli model) and to real data (genetic data on microsatellites on the Y

chromosomes of human populations from Central Asia, with their social and geographic structures).

ABC estimation of the ecological parameters based on the genealogical tree

The dataset denoted $\mathbf{z}$ contains the genotype and/or phenotype on the trait x and the marker u for each of the n sampled individuals. The trait x can be geographic locations, species or strain identity, size, color, genotypes or anything that affect the ecological parameters and fitness. The marker u can also be genotypic or phenotypic measures, discrete or continuous, qualitative or quantitative, but with no effect on fitness (the marker is supposed neutral). Our goal is to use the dataset $\mathbf{z}$ to estimate the parameters of the model denoted θ (in our case, birth and death rate, competition kernel, mutation probabilities and kernel) using an ABC approach. To do so, the following procedure is repeated a large number of times:

- 1st step. A parameter set θ_i is drawn in a prior distribution $\pi(d\theta)$;
- 2nd step. A PES and its neutral nested genealogies of the n sampled individuals are simulated in each model associated with the parameters θ_i ;
- 3rd step. A set of summary statistics S_i is computed from the data simulated under θ_i , for each i .

The posterior distribution of the model is then approximated by comparing, for each simulation i , the simulated summary statistics S_i to the ones from the real dataset and by computing for each parameter θ_i a weight W_i that defines the approximated posterior distribution (see Formula B.1 in supplementary materials). Three categories of summary statistics have been used, each associated with a different aspect of the genealogical tree (the complete list of summary statistics is given in Appendix D in supplementary materials):

- The trait distribution describing the strains diversity and their abundances (e.g., number of strains, the mean and variance of strains abundance, ...);
- The marker distribution in the sampled population describing the neutral diversity within each sampled strain (e.g., the M-index, F_{st} , Nei genetic distances,...);
- The shape of the genealogy (e.g., most recent common ancestor, length of external branches, number of cherries, ...).

Depending on the dataset and the information available for a given population, four scenarios can be encountered:

Scenario 1. Complete information: The evolutionary history of the trait and the genealogies, populations and

subpopulations abundances, values of the sampled individuals on the trait x and the marker u . This situation certainly never occurs but it is a reference which allows to evaluate the expected ABC estimation in a perfect situation where all information is available. This situation can also include cases where independent information can be added such as fossil records;

Scenario 2. Population information: Total population abundance, values of the trait x and marker u of the sampled individuals. The estimations given with those statistics represent the estimations one could expect with a complete knowledge of the present population;

Scenario 3. Sample information: The number of sampled sub-populations, the values of the trait x and the marker u of the sampled individuals;

Scenario 4. Partial sample information: Only the number of sampled sub-populations and the values of the marker u of the sampled individuals.

The four situations will be compared regarding the quality of the ABC estimations of the model parameters.

Application 1: Inference of the parameters in the Dieckmann–Doebeli model

In this section, we applied the ABC statistical procedure on the traits distribution and their phylogenies generated by a simple eco-evolutionary model (Champagnat et al. 2006, Dieckmann and Doebeli 1999, Roughgarden 1979). The birth rate of an individual with trait x is $b(x) = \exp(-x^2/2\sigma_b^2)$, the individual natural death rate is constant $d(x) = d_c$, and the competition between two individuals with traits x and y is $C(x, y) = \eta_c \exp(-(x - y)^2/2\sigma_c^2)$, $\sigma_c > 0$. The trait space is chosen to be $\mathcal{X} = [-1, 1]$. The effect of a mutation on the trait x is randomly drawn in a Gaussian mutation kernel with mean 0 and variance σ_m^2 (values outside $\mathcal{X}$ are excluded). The probability of mutation is p . The markers are assumed to be a vector of 10 microsatellites, each of them mutating with the same rate q . When a microsatellite mutates, we increase or decrease its value by 1 with equal probability.

The distribution of the phylogenies depends on the parameter $\theta = (p, q, \sigma_b, \sigma_c, d_c, \eta_c, t_{sim})$, where t_{sim} is the duration of the PES (t_{sim} is not known a priori and must be considered as a nuisance parameter).

Posterior distribution and parameters estimation

We ran $N = 400,000$ simulations with identical prior distributions and scaling parameter $K = 1000$ (see details in Appendix B). Chosen parameter sets and prior distributions are given in Appendix A.4. We randomly chose four simulations runs among the N simulations as pseudo datasets (these sets are named A, B, C, and D, see Appendix C, Table 1 and Fig. 1). All other simulations runs were used for

the parameters estimation. Figure 2 shows the posterior distribution for one of the pseudo dataset (see Appendix E for full results). Our results show that estimates based on all statistics (Scenario 1, blue distribution) are not always the most accurate, suggesting that some of the descriptive statistics introduce noise and worsen estimate accuracy. However, the descriptive statistics providing knowledge about how population is trait-structured do not belong to this group and importantly improve estimation when available (compare orange vs. red posterior distributions).

The impact of the number of microsatellites on the quality of the estimation is tested for the first pseudo dataset A (see Appendix C, Table 1) with the number of microsatellites varying from 10 to 100. A sensitivity analysis is shown in the supplementary materials, Fig. E.4: the results are quite robust to this number. For some parameters such as t_{sim} , better precision is achieved with increased number of microsatellites, and for other parameters such as q or p , the impact of the number of microsatellites is more visible under Scenario 4 when we should rely a lot on the information brought by the microsatellites.

Discrepancy with Kingman's coalescent

After a correct renormalization, Kingman's coalescent are generally considered as a good approximation of coalescent trees, even in structured populations. However, in our model, the population structure itself can evolve, demographic rates can vary with time, and subpopulations can interact with each other, which might strongly affect the topology of the coalescent trees and their branches length. In this section, our aim is to evaluate to what extent the Kingman's coalescent is a good approximation or not of the genealogies generated by the Doebeli–Dieckmann's model. In case of a significant discrepancy, we further determined the properties of the trees which show important differences between both models, and then we identified and evaluated the type and extent of errors that one would expect when using Kingman's coalescents for inference without taking into account the evolution of population structure.

We considered statistics commonly used to test the neutrality of the phylogenies of n sampled individuals (Fu and Li 1993): the number of cherries C_n , i.e., the number of internal nodes of the tree having two leaves as descendants, the length of external branches L_n , i.e., edges of the phylogenetic tree admitting one of the n leaves as extremity, and the time T_n^{MRCA} to the most recent common ancestor (MRCA). The distributions of the normalized C_n and L_n and the distribution of T_n^{MRCA} for the forward–backward Doebeli–Dieckmann's coalescent and the Kingman's coalescent are compared. For Kingman's coalescent, asymptotic normality has been established for C_n and L_n (see (Blum and François 2005, Janson and Kersting 2011)). The distribution of T_n^{MRCA} for the Kingman

Fig. 1 Dynamics of **a** the trait x and **b** the neutral marker u of the four pseudo datasets **A**, **B**, **C**, and **D**. These pseudo datasets are randomly sampled among $N = 400,000$ simulations runs of the Doebeli–Dieckmann’s model (Parameter sets are given in Appendix Table 1). Figures show the Substitution Fleming–Viot Process (SFVP) and the nested phylogenetic tree of n individuals sampled at the final time of the simulation. **a** The trait x follows a Polymorphic

Evolutionary Substitution (PES) process introduced in Champagnat and Méléard (2007). **b** The genealogies of the marker u follow a forward–backward coalescent process nested in the PES tree as described in section “Genealogies in a forward–backward coalescent with competitive interactions”. The colors refer to the lineage to which one individual belongs shown in **a**.

Fig. 2 Prior and posterior distributions (pseudo dataset A in Fig. 1). Black dashed curve: prior distribution; Vertical red line: true value. The different colors correspond to different scenario regarding which data are available or not: Blue, Scenario 1 (All descriptive statistics are

available); Pink, Scenario 2 (data from the totality of the population); Red, Scenario 3 (data from a sample of the population); Orange, Scenario 4 (data from a sample of the population, the traits x is not known). Results for other pseudo datasets are given in Appendix E.

Fig. 3 Testing the difference with a Kingman's coalescent. **a** External branch length L_n : Box-plot of the p values of the Kolmogorov–Smirnov test, for each value of the number of lineages m at sampling time (in abscissa). **b** Number of cherries C_n : Box-plot of the p values of the Kolmogorov–Smirnov test, as a function of m . For **a** and **b**, 100 ABC analysis were done for each value of m and we tested if the distribution of the normalized external branch length follows a Gaussian distribution (H_0). The threshold value of rejection

of H_0 , 0.05, is represented by the dashed red line. If the p values are lower than this threshold, the distribution of the statistics (L_n or C_n) of the forward–backward coalescent trees generated by a Doebeli–Dieckmann model is significantly different than the one under a Kingman's coalescent. **c** Compared distributions of the age of the MRCA for the forward–backward coalescent (plain line) and for the Kingman's coalescent (dotted line).

coalescent is computed by using the fact that the trees are binary with exponential durations between each coalescence. Neutrality tests conditionally on the number of lineages m at the time of sampling are performed using the behavior of these statistics under the null assumption H_0 that the phylogenies correspond to a Kingman's coalescent. For each m , we chose as pseudodata one of the simulations of our model with m species at the final time, and we performed normality tests for C_n and L_n , and an adequation test for the expected distribution under Kingman for T_n^{MRCA} . This was repeated 100 times for each value of $m \in \{1, \dots, 10\}$ (details given in Appendix F).

Figure 3 shows the distributions of the *a posteriori* p values for the normality tests for L_n and C_n . The coalescent trees significantly differ from Kingman's coalescent trees regarding the external branch length L_n (Fig. 3(a)), while the number of cherries C_n is not always significantly different (the p -values have a median close to 0.05, Fig. 3(b)). Finally, Fig. 3(c) shows the distribution of the time to the MRCA depending on the number of lineages m . A mean comparison test shows that the mean of the T_n^{MRCA} s obtained from the simulations of our forward–backward coalescent significantly differs from the expected MRCA time under a Kingman's coalescent (see Appendix F.2). Hence, our results show that coalescent tree topologies generated under a Doebeli–Dieckmann's model are expected to be significantly different from a Kingman's coalescent.

Figure 4 shows further comparison between Kingman's coalescent and the trees under our model. The distribution of external branch lengths under our model follows an asymmetrical leptokurtic distribution and it tends to be much shorter than under a Kingman's coalescent. The time to the MRCA is also much longer under our model than the Kingman's coalescent. The distribution of the number of cherries follows a symmetrical bell-shaped distribution flattened around the mode.

Overall, we found that the coalescent trees generated by a Doebeli–Dieckmann model significantly differ from a Kingman's coalescent. In particular, we found that using a Kingman's coalescent model and ignoring the trait structure of a population tend to overestimate the recent coalescent times. The genealogies generated by the forward–backward coalescent under a Doebeli–Dieckmann's model are expected to differ from a standard or renormalized Kingman's coalescent for various reasons: (i) there are multiple instantaneous coalescence events when a new lineage appears; (ii) coalescence rates differ among lineages, creating asymmetries in the phylogenetic tree (trees can therefore be imbalanced); (iii) coalescence rates vary in time since they depend on the structure of the population and the traits present at a given time; and (iv) eco-evolutionary feedbacks and competitive interactions between lineages affect coalescent rates in the whole population.

Fig. 4 Histograms of **a** the renormalized external branch lengths, **b** the renormalized number of cherries, **c** the time to the MRCA. The simulations are shown for $p = 0.0076$, $q = 0.7503$, $\sigma_b = 1.186$, $\sigma_c = 0.4951$, $\sigma_m = 0.1448$, $\eta_c = 0.0211$ and $t_{sim} = 1025.619$ (set of

parameter A in Table 1. Results for three other ‘reference’ sets are given in App. F). The dashed line represents the distribution followed by a Kingman’s coalescent (Gaussian distribution for **a** and **b**, simulations for **c**).

Application 2: correlations between genetic and social structures in Central Asia

In Anthropology, a common question is whether or not socio-cultural changes can affect demographic parameters, such as fertility rates. For instance, it is hypothesized that agriculturalists have a higher fertility than foragers (e.g., (Sellen and Mace 1997)), which is supported by several studies (e.g., (Bentley and Goldberg 1993, Ross et al. 2016)). In this section, we analyze genetic data in order to test whether populations with two different lifestyles and social organizations show different fertility rates. Nineteen human populations from Central Asia have been sampled in previous studies (Fig. 5(a), Chaix et al. (2007), Heyer et al. (2015)). Two types of socio-cultural organizations are encountered: Indo-iranian populations are patrilineal, i.e., mostly pastoral and organized into descent groups (tribes, clans...); Turkic populations are cognatic, i.e., mostly sedentary farmers organized in nuclear families. 631 individuals have been sampled (310 from a cognatic population, 321 from a patrilineal one). Ten microsatellite loci have been genotyped on the Y-chromosome. Since there is no recombination on the sexual chromosomes in humans, it is appropriate to use our model which assumes clonal reproduction. Hence, we will perform ABC analysis on the genetic diversity following the paternal lineages.

We considered that the trait x in the model is a vector containing the geographic location of the population and the social organization (cognatic or patrilineal). For geographical positions, given the Fig. 5(a), we consider that geographic location is 1-dimensional: we can fit a

polynomial curve through the geographical positions of the tribes:

$$P(x) = 673.4 - 25.13x + 0.327x^2 - 1.39 \cdot 10^{-3}x^3 \quad (R^2 = 0.92).$$

Hence the location of each population is given by the coordinates $(x, P(x))$ (Fig. 5(b)). The distance between populations is computed thanks to the line integral along the interpolated curve (see details in Appendix G.2). The neutral marker u is a vector containing the genotype at the ten microsatellites. Here we assume that the neutral marker is fully linked with the trait corresponding to the social organization.

Our aim is to use our ABC procedure on the genetic data to estimate the parameters $\theta = (p_{xb01}, b_0, b_1, p_{loc}, q, \sigma_{loc}, \eta_0, \eta_1, \sigma_c, t_{sim})$ of our model. The individual birth rates is assumed to depend on social organization only and not on geographic location: b_0 for the patrilineal populations and b_1 for the cognatic ones. Death rates are supposed to be due to density-dependent competition for the sake of simplicity: the competitive effect of an individual located at coordinate y on an individual in a patrilineal (resp. cognatic) population at location y' is supposed $C(y, y') = \eta_0 \exp(-(y - y')^2 / 2\sigma_c^2)$ (resp. $C(y, y') = \eta_1 \exp(-(y - y')^2 / 2\sigma_c^2)$). The individual death rate at location y is given by the sum of the competitive effects of all individuals. We supposed that, with probability p_{loc} , an individual can found a new population after dispersal (corresponding to a mutation on the trait x at birth, in other words we supposed for simplicity that each new population is

a

b

Fig. 5 Geographic locations of the studied human populations from Central Asia. **a** Map of sampling locations from Heyer et al. (2015). Triangles correspond to cognatic Indo-Iranian populations, quares to patrilineal Turkic populations. **b** Regression of the data to a 1-dimensional problem.

founded by a single individual). With probability p_{xb01} , a social organization change can occur. The location of the new population is randomly drawn in a centered Gaussian with standard deviation σ_{loc} . Following anthropological data, we assumed that social organization changes are unidirectional only from patrilineal pastoral to cognatic farmers populations (Chaix et al. 2007). t_{sim} and q respectively are the duration of the coalescent and the marker mutation probability.

Estimating the parameter θ and using the ABC procedure to select between alternative models will allow us to test whether the null hypothesis

$$H_0 : b_0 = b_1, \quad (3.1)$$

is acceptable, compared to the alternative hypothesis $H_a: b_0 < b_1$ (see e.g., (Grelaud et al. 2009, Prangle et al. 2013, Stoeck et al. 2015)). We generated a set of data with the a

Fig. 6 Results of the ABC estimation for the dataset of Heyer et al. (2015) for Central Asia human populations. The prior distributions are plotted in dashed lines and the posterior densities in plain red lines.

priori probability $1/2$ of having $b_0 = b_1$ and the a priori probability $1/2$ of having $b_0 < b_1$ (see details in Appendix G.2). To do this, we generated 10,000 datasets with $b_0 = b_1$ and 10,000 datasets with $b_0 < b_1$. The ABC estimation provides weights W_i for each of these 20,000 simulations (see Eq. B.1) yielding the posterior distribution of the parameters (see Fig. 6 and 7). These weights W_i also allow to compute the posterior probabilities of each hypothesis, H_0 : $\{b_0 = b_1\}$ or H_a : $\{b_0 < b_1\}$. When the estimated posterior probability for $\{b_0 < b_1\}$ is larger than a certain threshold α , the null hypothesis H_0 is rejected.

We first checked the quality of the ABC estimation and of the test (3.1) on simulated data. Among the 20,000 simulations presented in the above paragraph, we chose 200 simulations to play in turn the role of the true dataset, 100 among those with $b_0 = b_1$ and 100 among those with $b_0 < b_1$. We obtained that parameters estimates were generally close to the true values (Appendix G.2 in the supplementary materials). We then use these 200 datasets to perform 200 tests (using for each of them the 19,999 other simulations). Since we know for each of these 200 tests whether the data are obtained under H_0 : $\{b_0 = b_1\}$ or H_a : $\{b_0 < b_1\}$, this provides insight on the power of our test and allows us to set the threshold defining the critical region of the test. Here we can choose this threshold $\alpha = 0.5$ which is very natural (see Appendix G.2). We can then conclude the test for the dataset from Central Asia populations.

For the ABC test, we obtained an estimated posterior probability for $\{b_0 < b_1\}$ equal to 0.4518, below the threshold $\alpha = 0.5$, so that the null hypothesis H_0 (3.1) can not be rejected. The p -value of the test, estimated as the proportion of these simulations where $\hat{\mathbb{P}}(H_a | S_{obs}) \geq 0.4518$, can be estimated to 47%. Hence

there is no significantly higher fecundity in cognatic populations compared with patrilineal ones.

Discussion

Inferences from genetic data are most often performed under three important assumptions in the existing literature. First, the population size and structure are known parameters: either it is fixed or it follows a deterministic evolution, according to a given scenario (e.g., expansion or bottleneck, or a fixed structure with known migration rates between sub-populations). Second, mutations are supposed to not affect the genealogical trees, i.e., models are supposed neutral. Selection is rarely explicitly taken into account in inference methods (yet see for instance (Charlesworth et al. 2003, Johri et al. 2020), where background selection can bias the estimation of demographic variations). Third, there is no feedback between the evolution of the population and its demography: a selected mutation is supposed not to affect the population size, or the population structure. The most frequent models used in inference, the Kingman's coalescent and the Wright–Fisher model, make the three assumptions altogether. The goal of the present paper was to present a model and an inference method which allow to relax all these assumptions. We showed that by using an ABC procedure, it was possible to estimate ecological, demographic and genetic parameters from genotypic and phenotypic data.

Recently, Rasmussen and Stadler (2019) proposed a birth-death model without interactions where mutations can affect the birth and death rates of individuals in a strain, which in return affects the genealogies. They showed how it

Fig. 7 Approximate posterior distributions for b_0 , b_1 and $b_1 - b_0$. **a** Approximate posterior distributions for b_0 and b_1 , obtained by ABC on the Central Asian database with 40,000 simulations. **b** Approximate posterior distribution of $b_1 - b_0$. The posterior mean of $b_1 - b_0$, equal to 0.25, is indicated as the vertical dashed red line.

was possible to use phylogenies to estimate the effect of mutations on fitness in some viruses. In our paper, we go a step further by allowing interactions between individuals, and population structure and demography that depend on the evolution of the population. Our model assumes two genetics traits, a selected trait which governs the structure of the population, and a marker linked to the trait which is neutral and used to infer the genealogy. We first showed how genetic diversity at the neutral marker is related to the evolution at the selected trait, and to the size and structure of the population. We then used this relationship by developing an ABC procedure which allows to estimate ecological parameters based on genetic diversity at the neutral marker and on the partial or total knowledge of the population structure. We showed on simulated data that the ABC procedure gives accurate estimates of ecological parameters such as the birth, death and interactions rates, and genetic parameters such as the mutation rate. Our results also showed that non-neutral genealogies can easily be detected under our framework.

The ABC procedure is well fitted to deal with complex models if we can simulate the latter easily, which has become increasingly common for most ecological models (e.g., (Haller and Messer 2019, Legendre et al. 1995)).

Here, we applied our model and its ABC procedure to reanalyze the genetic diversity of microsatellites on Y chromosomes in Central Asia human populations. Genetic diversity is compared between two social organizations and lifestyles: patrilineal vs. cognatic. Previous studies showed significantly different genetic diversity and coalescent trees topologies, which was interpreted as evidence of the effect of sociocultural traits on biological reproduction, due to how wealth is transmitted within families (Chaix et al. 2007, Heyer et al. 2015). However, these conclusions were obtained under simplifying assumptions: genealogies followed a modified Wright-Fisher model, and the genetic diversity and coalescent trees topologies were compared independently, i.e., there was no interaction between populations and between social organization. Such assumptions dismissed the possibility that sociocultural traits and social organization could change, that new populations can be founded, and that competitive interactions between individuals within and between social organizations might affect demography and evolution. We relaxed all these limitations by applying our model. We supposed that the trait under selection can affect the birth rate. Contrarily to Heyer et al. (2015), we did not test whether wealth transmission could explain differences in

genetic diversity and coalescent trees topologies. Rather, we addressed a long-standing question in anthropology: can fertility be affected by a change in a social organization, in particular with a change in the agricultural mode. We found no evidence of a fertility difference between both kinds of social organization. Our findings then ask the question why human populations can adopt new sociocultural traits without any strong evidence of a biological advantage. Further analyses and data would be necessary to confirm our results, especially regarding the number of children per females. In the data, this information is based on a few interviews that lack precision (see Table S3 in the supplementary material of (Chaix et al. 2007)). However, since the genetic diversity sampled in contemporaneous population is due to long historical process, it seems difficult to estimate fertility for several dozens or hundreds generations. Our results only suggest that there is, on average, no evidence of an effect of a social trait on fertility all along the history of Central Asia human populations.

Our results on the reanalysis of sociodemographic parameters in Central Asia human populations should however be taken cautiously because the posterior distributions of some parameters were not narrow enough. We believe that this limited accuracy is due to several factors. First, the models we chose are complex and many parameters are estimated, which has an inferential cost. This is a general difficulty for inferential methods from genetic data, particularly in our case since our aim is to estimate at the same time genetic, ecological and demographic parameters. Determining the extent to which we can expect accuracy of parameters estimation in such complex models is still an open and challenging question. In any case, it is actually difficult to assess the quality of our inferences without alternative methods and models addressing the same questions. Second, our estimations are based on limited genetic information (a dozen microsatellites). We expect our estimations to be largely improved with more genetic and genomic data. Further developments of our methods to such type of data, for instance SNPs, are yet needed before analyzing other datasets. We also want to stress out that datasets containing both genomic, ecological and phenotypic data are scarce, which would actually limit our capacity to apply our method on other datasets. We hope that the development of methods such as ours will motivate the collection of more integrated data in the future.

Our model is based on classical competitive Lotka–Volterra equations, under the assumptions of rare mutations relatively to ecological processes. The genealogies and genetic diversity produced under such a model are then used to infer ecological and demographic parameters. We showed that relaxing strong assumptions of genetic models is possible, and that it allows to provide new analysis methods based on the ABC procedure. Even though we applied our inferential procedure only to

simulated genetic data or microsatellites genetic diversity, our model is general enough to embrace any type of data: SNPs, phenotypic traits, etc. The development of stochastic birth and death models, with (this paper) or without (Rasmussen and Stadler 2019) interactions open the way to new methods for analyzing data. As highlighted by Frost et al. (2015), this is particularly important for the study of epidemics and pathogens evolution. These authors give a list of current challenges which can partly be addressed thanks to the method and models developed here. For instance, studying the role of the host structure on the pathogens evolution and genetic diversity, or the role of stochasticity, can be done along these lines using more complex and realistic evolutionary models.

Data availability

The summary statistics from genetic data, simulations programs and scripts for the analysis are available at https://github.com/ClotildeLepers/Coalescent_process_ABC_application

Acknowledgements The authors thank Laurent Séries and Sylvain Ferland for their help with the CMAP compute servers. They also thank Frédéric Austerlitz and Raphaëlle Chaix for discussion and for sharing anthropological data from Central Asia. This research has been supported by the Chair "Modélisation Mathématique et Biodiversité" of Veolia Environnement-Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X. VCT also acknowledges support from Labex CEMPI (ANR-11-LABX-0007-01) and Bézout (ANR-10-LABX-58).

Compliance with ethical standards

Competing interests The authors declare no competing financial interests in relation to the current work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Barton N (1998) The effect of hitchhiking on neutral genealogies. *Genet Res* 72:123–133
- Barton N (2000) Genetic hitchhiking. *Phil Trans R Soc Lond B* 355:1553–1562
- Barton NH, Etheridge AM, Véber A (2010) A new model for evolution in a spatial continuum. *Electron J Probab* 15:162–216
- Beaumont M, Cornuet JM, Marin JM, Robert C (2009) Adaptive approximate Bayesian computation. *Biometrika* 96:983–990
- Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035
- Bedford T, Cobey S, Pascual M (2011) Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol Biol* 11:220
- Beichman A, Huerta-Sanchez E, Lohmuller K (2018) Using genomic data to infer historic population dynamics of nonmodel organisms. *Annu Rev Ecol, Evol Syst* 49:433–456
- Blancas, AB, Duchamps, JJ, Lambert, A, Siri-Jégousse, A (2018) Trees within trees: simple nested coalescents. *Electron J Probab*. 23:1–27

- Blancas, AB, Gufler, S, Kliem, S, Tran, V, Wakolbinger, A (2020) Evolving genealogies for branching populations under selection and competition. In preparation
- Bentley G, Goldberg T, ska G (1993) The fertility of agricultural and non-agricultural traditional societies. *Popul Stud* 47:269–281
- Billiard S, Ferrière R, Méléard S, Tran V (2015) Stochastic dynamics of adaptive trait and neutral marker driven by eco-evolutionary feedbacks. *J Math Biol* 71:1211–1242
- Blum M (2010) Approximate Bayesian Computation: a non-parametric perspective. *J Am Stat Assoc* 105:1178–1187
- Blum M, François O (2005) On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Math Biosci* 195:141–153
- Blum M, François O (2010) Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* 20:63–73
- Blum M, Tran V (2010) HIV with contact-tracing: a case study in Approximate Bayesian Computation. *Biostatistics* 11:644–660
- Chaix R, Quintana-Murci L, Hegay T, Hammer M, Mobasher Z, Austerlitz F et al. (2007) From social to genetic structures in Central Asia. *Curr Biol* 17:43–48
- Champagnat N (2006) A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stoch Processes Appl* 116:1127–1160
- Champagnat N, Ferrière R, Méléard S (2006) Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models via timescale separation. *Theor Popul Biol* 69:297–321
- Champagnat N, Jabin PE, Raoul G (2010) Convergence to equilibrium in competitive Lotka-Volterra and chemostat systems. *Comptes-Rendus Mathématiques de l'Académie des Sciences de Paris* 348:1267–1272
- Champagnat N, Méléard S (2007) Invasion and adaptive evolution for individual-based spatially structured populations. *J Math Biol* 55:147–188
- Champagnat N, Méléard S (2011) Polymorphic evolution sequence and evolutionary branching. *Probab Theory Relat Fields* 151:45–94
- Charlesworth B, Charlesworth D, Barton NH (2003) The effects of genetic and geographic structure on neutral variation. *Annu Rev Ecol Evol Syst* 34:99–125
- Csillery, K, Francois, O, Blum, MGB (2012) abc: an r package for Approximate Bayesian Computation (ABC). *Methods in ecology and evolution*.
- Dawson D, Hochberg K (1982) Wandering random measures in the Fleming-Viot model. *Ann Probab* 10:554–580
- Dawson, DA (1993) Measure-valued Markov processes. In: Springer (ed.), *Ecole d'Eté de probabilités de Saint-Flour XXI*. New York, *Lectures Notes in Math*, 1541, p. 1–260
- Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature* 400:354–357
- Donnelly P, Kurtz T (1996) A countable representation of the Fleming-Viot measure-valued diffusion. *Ann Probab* 24:698–742
- Donnelly P, Kurtz T (1999) Particle representations for measure-valued population models. *Ann Probab* 27:166–205
- Duchamps, JJ (2020) Trees within trees ii: nested fragmentations. *Ann. Inst. H. Poincaré Probab. Statist.* 56:1203–1229
- Durrett R, Schweinsberg J (2004) Approximating selective sweeps. *Theoret Popul Biol* 66:129–138
- Durrett R, Schweinsberg J (2005) Random partitions approximating the coalescence of lineages during a selective sweep. *Ann Appl Probab* 15:1591–1651
- Etheridge, A (2000) *An introduction to superprocesses*, University Lecture Series, vol. 20, American Mathematical Society, Providence.
- Etheridge A, Pfaffelhuber P, Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. *Ann Appl Probab* 16:685–729
- Ethier S, Kurtz T (1986) *Markov processus, characterization and convergence*. John Wiley, Sons, New York
- Felsenstein J (1975) A pain in the torus: some difficulties with the model of isolation by distance. *Am Nat* 109:359–368
- Fournier N, Méléard S (2004) A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *Ann Appl Probab* 14:1880–1919
- Frost S, Pybus O, Gog J, Viboud C, Bonhoeffer S, Bedford T (2015) Eight challenges in phylodynamic inference. *Epidemics* 10:88–92
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709
- Gallieni L (2017) Intransitive competition and its effects on community functional diversity. *Oikos* 126:615–623
- Goldstein D, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. *Annu Rev Genom Hum Genet* 3:129–152
- Grelaud A, Robert C, Marin J, Rodolphe F, Taly J (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal* 4:317–336
- Haller BC, Messer PW (2019) SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol* 36(3):632–637
- Heyer E, Brandenburg JT, Leonardi M, Toupance B, Balaesque P, Hegay T et al. (2015) Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *Am J Phys Anthropol* 157:537–543
- Jansen S, Kurt N (2014) On the notion(s) of duality for Markov processes. *Probab Surveys* 11:59–120
- Janson S, Kersting G (2011) On the total external length of the kingman coalescent. *Electron J Probab* 16:2203–2218
- Johri, P, Riall, K, Jensen, JD (2020) The impact of purifying and background selection on the inference of population history: problems and prospects. *bioRxiv*. <https://doi.org/10.1101/2020.04.28.066365>
- Legendre S, Clobert J (1995) ULM, a software for conservation and evolutionary biologists. *J Appl Stat* 22:817–834
- Marin JM, Pudlo P, Robert C, Ryder R (2012) Approximate Bayesian computation methods. *Stat Comput* 22:1167–1180
- Metz J, Geritz S, Meszéna G, Jacobs F, Heerwaarden JV (1996) Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. In: Van Strien SJ, Verduyn Lunel SM (eds) *Stochastic and spatial structures of dynamical systems*. pp 183–231
- Müller N, Rasmussen DA, Stadler T (2017) The structured coalescent and its approximations. *Mol Biol Evol* 34:2970–2981
- Neher R, Bedford T (2015) nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* 31:3546–8
- Pitman J (1999) Coalescents with multiple collisions. *Ann Probab* 27:1870–1902
- Prangle, D, Fearnhead, P, Cox, M, Biggs, P, French, N (2013) Semi-automatic selection of summary statistics for abc model choice. *Stat Appl Genet Mol Biol* pp. 1–16
- Pudlo P, Marin J, Estoup A, Cornuet J, Gautier M, Robert C (2016) Reliable ABC model choice via random forests. *Bioinformatics* 32:859–866
- Rasmussen D, Stadler T (2019) Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *eLife* 8:e45562
- Ross CT, Mulder MB, Winterhalder B, Uehara R, Headland J, Headland T (2016) Evidence for quantity-quality trade-offs, sex-specific parental investment, and variance compensation in colonized agta foragers undergoing demographic transition. *Evol Hum Behav* 37:350–365
- Roughgarden J (1979) *Theory of population genetics and evolutionary ecology: an introduction*. Macmillan, New York

- Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. *J Appl Probab* 36:1116–1125
- Sellen D, Mace R (1997) Fertility and mode of subsistence: a phylogenetic analysis. *Curr Anthropol* 38:878–889
- Spor A, Nidelet T, Simon J, Bourgaïs A, de Vienne D, Sicard D (2009) Niche-driven evolution of metabolic and life-history strategies in natural and domesticated populations of *Saccharomyces cerevisiae*. *BMC Evol Biol* 9:296
- Stephan W (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol Ecol* 25:79–88
- Stoehr J, Pudlo P, Cucala L (2015) Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields. *Stat Comput* 25:129–141
- Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics* 192:671–682
- Verdu P, Austerlitz F, Estoup A, Vitalis R, Georges M, Théry S et al. (2009) Origins and genetic diversity of pygmy hunter-gatherers from western central africa. *Curr Biol* 19:312–318
- Zeeman M (1993) Hopf bifurcations in competitive three-dimensional Lotka-Volterra systems. *Dynam Stab Syst* 8:189–217