Approximate Bayesian Computation applied to time series of population genetic data disentangles rapid genetic changes and demographic variations in a pathogen population

Méline Saubin, Aurélien Tellier, Solenn Stoeckel, Axelle Andrieux, Fabien Halkett

Approximate Bayesian Computation applied to time series of population genetic data disentangles rapid genetic changes and demographic variations in a pathogen population

Méline Saubin[1,2], Aurélien Tellier[2], Solenn Stoeckel[3], Axelle Andrieux[1], Fabien Halkett[1]

[1] Université de Lorraine, INRAE, IAM, F-54000 Nancy, France

[2] Professorship for Population Genetics, Department for Life Science Systems, Technical University of Munich, Freising, Germany

[3] INRAE, Agrocampus Ouest, Université de Rennes, IGEPP, F-35653 Le Rheu, France

Corresponding author:

Fabien Halkett

INRAE Centre Grand-Est – Nancy, UMR 1136 Interactions

Arbres-Microorganismes, F-54280, Champenoux, France

E-mail: fabien.halkett@inrae.fr

Abstract

Adaptation can occur at remarkably short timescales in natural populations, leading to drastic changes in phenotypes and genotype frequencies over a few generations only. The inference of demographic parameters can allow understanding how evolutionary forces interact and shape the genetic trajectories of populations during rapid adaptation. Here we propose a new Approximate Bayesian Computation (ABC) framework that couples a forward and individual-based model with temporal genetic data to disentangle genetic changes and demographic variations in a case of rapid adaptation. We test the accuracy of our inferential framework and evaluate the benefit of considering a dense versus sparse sampling. Theoretical investigations demonstrate high accuracy in both model and parameter estimations, even if a strong thinning is applied to time series data. Then, we apply our ABC inferential framework to empirical data describing the population genetic changes of the poplar rust pathogen following a major event of resistance overcoming. We successfully estimate key demographic and genetic parameters, including the proportion of resistant hosts deployed in the landscape and the level of standing genetic variation from which selection occurred. Inferred values are in accordance with our empirical knowledge of this biological system. This new inferential framework, which contrasts with coalescent-based ABC analyses, is promising for a better understanding of evolutionary trajectories of populations subjected to rapid adaptation.

1 Introduction

Adaptation can occur at remarkably short timescales in natural populations, leading to drastic changes in genotype frequencies and phenotypes over a few generations only (Buffalo and Coop, 2019). This rapid pace of adaptation has motivated the use of temporal data to understand neutral (Prout, 1954; Wallace, 1956; Nei and Tajima, 1981; Pollak, 1983; Mueller et al., 1985b; Waples, 1989; Wang and Whitlock, 2003) and under-selection genetic evolution over time (Dobzhansky, 1943; Dobzhansky and Pavlovsky, 1971; Fisher and Ford, 1947; Kettlewell, 1958, 1961; Mueller et al., 1985a). However, such time series studies remain rare compared to the amount of work focusing on one contemporary sample to trace back its genetic history (Buffalo and Coop, 2019, 2020; Pavinato et al., 2022). Temporal data allow to track the changes in allele frequency through time, and therefore lead to a better understanding of evolutionary processes (Dehasque et al., 2020; Feder et al., 2021; Saubin et al., 2023b). In cases of rapid adaptation especially, the resulting genetic changes may be transient (Day and Gandon, 2007; Parsons et al., 2018) and require time sampling around the selection event to be highlighted (Saubin et al., 2022).

The inference of demographic parameters can allow understanding how evolutionary forces (especially genetic drift and selection) interact and shape the genetic trajectories of populations during rapid adaptation (Bergland et al., 2014; Živković et al., 2019). However, the difficulties in obtaining the likelihood of models including both demography, selection, and genetic drift (Pavinato et al., 2022; Luqman et al., 2021) lead to consider alternative approaches relying on simulations (Bazin et al., 2010; Laval et al., 2019). With the advent of computational approaches, Approximate Bayesian Computation (ABC) has become a standard approach for genetic analyses aiming at tracing back the evolutionary history of populations (Rosenberg and Nordborg, 2002; Bazin et al., 2010; Cornuet et al., 2010; Estoup and Guillemaud, 2010; Collin et al., 2021). These methods, coupled with a coalescent simulator, allow inferring evolutionary scenarios and demographic parameters, by modelling the genealogy of the samples (Kingman, 1982; Whitlock and Barton, 1997). The coalescent is computationally very efficient but relies on strong modelling assumptions (Rosenberg and Nordborg, 2002). As such, coalescent-based ABC approaches suffer from two limitations. First, there is a gap between the models developed so far and the actual complexity of biological scenarios. Different coalescent

models have been developed to account for fluctuating population sizes (Sjödin et al., 2005) or partially clonal species (Orive, 1993; Ceplitis, 2003; Hartfield, 2021). Yet these specificities has been considered in isolation. To date, few models integrate both rapid demographic fluctuations and complex life cycles (Tellier and Lemaire, 2014). Second, coalescent-based inference methods are powerful only when coalescent events occur at the same time scale as the process being considered, such as past demographic events. This means that demographic events happening in a recent past compared to the coalescent time scale (in units of the effective population size N_e) cannot be inferred. For example, under antagonistic coevolution, changes in the demography of host and pathogen are too fast for a coalescent analysis, but tracking allele frequencies and nucleotide diversity forward-in-time is informative about this process (Živković et al., 2019). Rapid adaptation processes are associated with transient genetic changes that can be well illuminated by considering step-by-step modelling algorithms forward-in-time. The study of the rapid adaptation of pathogens would therefore benefit from being considered with forward models (Foll et al., 2015).

The ultimate goal of studying adaptation is to search for loci under selection in order to decipher the genetic architecture underlying adaptive events (Oleksyk et al., 2010; Hoban et al., 2016). In this regard, rapid adaptation poses a challenge because complex demography can lead to peculiar patterns of allele frequency changes, even at neutral loci. This in turn can blur the detection of selected loci, such as observed for complex population structures (De Mita et al., 2013). However, models to search for selected loci have so far considered relatively simple demography (e.g. Vitalis et al., 2014). When addressing the phenomenon of rapid adaptation, it may therefore be worthwhile to proceed in two steps (Luqman et al., 2021). The first step is to decipher how a rapid adaptation event shapes the population structure at neutral loci (Saubin et al., 2022) and to infer demographic parameter from these genetic changes. With a few exceptions (*e.g.* De Mita and Siol, 2012; Foll et al., 2015), are not considered in such frameworks. The second step is to perform accurate genome scan that explicitly takes into account the inferred demography (Luqman et al., 2021). Here we propose an original approach to infer demographic and ecological parameters from the rapid changes in allele frequency at neutral loci. We base our analyses on a new ABC framework that couples a forward and individual-based model with temporal genetic data to disentangle genetic changes and demographic variations

in a case of rapid adaptation.

Understanding and inferring the evolutionary trajectories of populations is of major interest to evolutionary biologists but it can also have practical applications for population management. This is especially the case in agriculture where the need to control pathogen populations is paramount. Pathogens can induce disease outbreaks devastating human-managed ecosystems (Anderson et al., 2004; Tobin, 2015; Savary et al., 2019). Understanding the evolution of pathogens is therefore crucial for developing effective disease management strategies (Bonneaud and Longdon, 2020). The rapid adaptation of pathogen populations (McDonald and Linde, 2002; Saubin et al., 2021) and the high stochasticity in pathogen evolutionary trajectories (Parsons et al., 2018) make this endeavour extremely difficult.

The rapid adaptation of pathogen populations in agrosystems results from the tremendous selection pressures exerted by modern agricultural practices (Zhan et al., 2015; Stukenbrock and McDonald, 2008). To counter plant disease outbreaks, breeders develop resistant plant genotypes. These genetic resistances are most often deployed across large spatial scales (Zhan et al., 2015; Rimbaud et al., 2021). These agricultural practices break eco-evolutionary feedbacks that maintain the polymorphism observed in natural host-parasite systems (Stukenbrock and McDonald, 2008; Brown and Tellier, 2011). As such, these practices weaken the sustainability of plant genetic resistances in favouring the emergence and spread of virulent (*i.e.* resistance-adapted) pathogens (Rimbaud et al., 2021; Saubin et al., 2023b).

Therefore, the outcome of plant genetic resistance deployment is often a resistance overcoming event, *i.e.* the failure of the host plant to remain resistant to the pathogen. This results in the spread of virulent pathogens on resistant hosts (Johnson, 1984; Pink and Puddephat, 1999; Brown and Tellier, 2011; Burdon et al., 2016). On the pathogen side, an event of resistance overcoming can translate into a strong selective sweep with the intense and unidirectional selection causing drastic demographic changes for the pathogen population (Burdon et al., 2016; Persoons et al., 2017; Saubin et al., 2021). Such rapid adaptation can lead to specific temporal genetic signatures at neutral loci, depending on the evolutionary scenario ruling the change in population sizes (Saubin et al., 2022).

Here, we propose to use ABC based on temporal genetic evolution to unravel the evolutionary scenarios

following rapid and contemporaneous adaptation. We apply our inferential framework to time series data and evaluate the added value of considering the full trajectory following forward simulations compared to few time samples. We test the accuracy of several temporal sampling designs in inferring model parameter values. Last, we apply our ABC inferential framework to empirical data describing the population genetic changes of the poplar rust pathogen following a major event of resistance overcoming (Persoons et al., 2017; Louet et al., 2023).

2 Materials and methods

2.1 Simulation model

The rapid adaptation we model is a resistance overcoming event that is monitored through time. We use an individual-based, forward-time and non-spatial demogenetic model, designed for diploid individuals (Saubin et al., 2021, 2022). This model couples population dynamics and population genetics to follow the evolutionary trajectory of different genotypes at the selected locus and at neutral genetic markers. The model is implemented in Python (version 3.7, van Rossum, 1995) and Numpy (Harris et al., 2020). We consider life cycles commonly found in temperate pathogen species, with seasonal variation in reproductive mode. These pathogens switch from clonal reproduction during the epidemic phase to sexual reproduction occurring once a year, in winter (Agrios, 2005). This general life cycle is adjusted in two variations to take into account that the sexual reproduction takes place either on the same host plant as for the epidemic phase or on an alternate host (usually a different species). These life cycles are named 'with' or 'without' host alternation, respectively (Boolean parameter *Cycle*). The life cycle of plant pathogens is generally well documented (Agrios, 2005). This is the case for the focal study species: the poplar rust fungus, *Melampsora larici-populina* (Basidiomycota, Pucciniales), (Pinon and Frey, 2005). But in some cases the full life cycle has only recently been elucidated (*e.g.*, for wheat stripe rust, *Puccinia striiformis*, Jin et al., 2010), or is still unresolved (*e.g.*, for coffee leaf rust, *Hemileia vastatrix*, Talhinhas et al., 2017). 'Without' host alternation, the model represents the evolution in time of a population of pathogens on two static host compartments: a resistant compartment

(R) and a susceptible compartment (S). Each compartment has a fixed carrying capacity for the pathogen population, K_R and K_S, respectively. We define two parameters: the cumulative carrying capacity of S and R ($K = K_R + K_S$); and the proportional size of R ($propR = \frac{K_R}{K}$). 'With' host alternation, the alternate host compartment (A) is added, to account for the sexual reproduction occurring on another species. This static compartment is assumed to be larger than the two other compartments, with a fixed carrying capacity K_A. In the following, we refer to the three compartments as S (susceptible hosts), R (resistant hosts), and A (alternate hosts).

We consider that a year consists of 11 generations, 10 rounds of clonal multiplication plus one sexual reproduction event (Hacquard et al., 2011). Three basic steps are modelled at each clonal generation: reproduction following a logistic growth (with growth rate (r) and carrying capacity K_R or K_S depending on the compartment considered), mutation of neutral loci (see below), and a two-way migration (migration rate mig), from S to R and vice versa (Figure 1, model specifications are detailed in Saubin et al. (2021)). At the end of clonal multiplication, random mortality is applied to the pathogen population (at rate τ) because some individuals fail to overwinter. Then, sexual reproduction occurs. It differs between life cycles, considering or not the obligate migration to the alternate host before mating. For the life cycle 'with' host alternation, the generation of sexual reproduction is followed by one generation of clonal multiplication on the alternate host before the obligate emigration to S and R. We simulate the evolution at neutral loci and at a selected locus responsible for the virulence (qualitative trait) of pathogen individuals. In accordance with the gene-for-gene model that is prevailing in host-pathogen interactions (Flor, 1971), virulence is a qualitative trait determined by a single locus and a recessive allele (named avr). Only virulent individuals (i.e. homozygous avr/avr at the virulence locus) can infect the resistant plants and invade R. We assume no fitness cost of virulence, that is all pathogen individuals (irrespective of their genotype at the virulence locus) have equal fitness on S. We do not consider mutation at the virulence locus, meaning that evolution stems only from standing genetic variation, with initial frequency f_{avr} of virulent allele introduced after the burn-in period. Evolution at neutral loci is set to suit classical population genetic analyses based on microsatellite markers: 23 loci with a maximum of 20 allelic states. For this model to be applicable to SNP markers as well, the mutational process

follows a k-allele model. The mutation rate is a particularly difficult parameter to estimate by ABC analyses (see for example Parat et al., 2016). However, mutation rate has a limited impact on the dynamics of genetic diversity compared to demographic variations and selection. It takes a large time scale to account for a small effect, which can be overlooked in the face of genetic drift, especially in partially clonal populations (Reichel et al., 2016). Therefore, for this analysis, we set the mutation rate of microsatellite markers to a fixed and realistic value of 10^{-3} (Ellegren, 2004).

Each simulation starts with genotypes randomly drawn from the 20 possible alleles followed by a burn-in period with a population of constant size $N = K_S$. During the burn-in period, the pathogen population only evolves on S. To reach the genetic drift and mutation equilibrium from the initial state, the burn-in period is set to $2N$ generations. For a diploid organism with frequent sexual reproduction, $2N$ generations are sufficient to reach a steady state (Reichel et al., 2016; Hartfield, 2021; Bessho and Otto, 2022). We build a random simulation design of 150,000 simulations, with input parameter values drawn randomly from defined prior distributions (Table 1).

Each simulation is run for 400 generations (with 11 generations per year), which amounts to 36 years. Simulation is aborted if the virulent allele goes extinct. Of the 150,000 simulations produced from the random simulation design, 106,058 lead to resistance overcoming and are used for the following analyses (50,547 'with' host alternation, and 55,511 'without' host alternation).

2.2 Summary statistics

Sampling represents a random draw of n individuals from a host compartment, with the sample size $n = 30$. The sampling occurs each year at the end of generation 9 on S and R, and additionally - for the life cycle 'with' host alternation - at generation 11 on A before the redistribution to S and R (Figure 1). Sampled individuals are not removed (sampling with replacement) so that their genotypes could contribute to subsequent generations.

To summarise the observed and simulated data sets, we compute the following statistics for each population (time sample in a given compartment): (i) the proportion of virulent individuals (P_{Vir}); (ii) the genotypic

diversity estimated by Pareto's β (β_P); (iii) the proportion of unique multilocus genotypes ($R = \frac{G-1}{n-1}$ with G the number of unique genotypes and n the number of sampled individuals); (iv) the genetic diversity estimated by Simpson's index (S); (v) the genetic diversity estimated by Shannon-Wiener's index (SW); (vi) the mean expected heterozygosity over all loci (MH_E); (vii) the variance of the expected heterozygosity over all loci (VH_E); (viii) the mean number of alleles per locus (MLA); (ix) the multilocus index of linkage disequilibrium ($\bar{r}_D$); the population differentiation index between populations on R and S sampled at the same generation ($F_{ST}R - S$); (x) the population differentiation index between the initial population (on S after burn-in) and the sampled population (TF_{ST}) (Table 2). All statistics are implemented along with the model. Summary statistics are calculated for each population, independently of the sampled compartment (S, R or A), except for P_{Vir} which is not recorded for populations sampled on R (because all individuals living on the resistant host are virulent), and for $F_{ST}R - S$ which is calculated from populations on S and R from the same generation. In the following analyses, summary statistics from different compartments are considered as distinct summary statistics. Because of the temporal sampling, each summary statistic is recorded for multiple generations, which constitutes a time series of summary statistics.

Two methods are used to perform the analyses: 1) Summary statistics from different generations are considered as distinct summary statistics and hence all generations along the time series are taken into account (hereafter referred to as 'Complete summary statistics'); 2) We extract the mean, variance, minimum and maximum values along each time series, and considered these four values as the distinct set of summary statistics for each population genetic index (hereafter referred to as 'Wrap-up summary statistics'). The total number of summary statistics used for each analysis is presented in Table S1.

2.3 Approximate Bayesian Computation

We use the `R` package `abc` (Csilléry et al., 2012) to perform an ABC analysis from the simulated data. This analysis involves two steps: model choice and parameter estimation. The model choice aims to distinguish between the two simulated life cycles, 'with' and 'without' host alternation. The parameter estimation aims to estimate each of the six quantitative parameters of the chosen simulation model (Table 1).

The model choice procedure is based on a weighted multinomial logistic regression (Fagundes et al., 2007) computed on 5% of the simulated data sets for which ϵ, the Euclidean distance between the summary statistics of the observed data set and the simulated data sets, is the smallest (Beaumont et al., 2002). Bayes factors are calculated as the ratio of the posterior probabilities for the tested models (Kass and Raftery, 1995). Posterior model probabilities are corrected for the number of simulations performed for each model, as implemented in the `postpr` function of the `abc` `R` package.

For the parameter estimation procedure, we use the 50,547 simulations obtained under the model 'with' host alternation. We estimate posterior distributions (mode and credible intervals calculated as 95% percentile intervals from the posterior distributions) of each parameter by applying the neural networks regression method (Blum and François, 2010) implemented in the `R` package `abc` based on the 5% of simulations closest to the observed data.

Cross-validation on simulated data

To assess the validity of the method and the identifiability of model parameters, cross-validations are performed at both steps of the analysis. Here we use identifiability in the broad sense, that is including quantitative differences in the ability to infer parameter values. A leave-one-out cross-validation for the model choice is performed on 500 randomly drawn simulations from each of the two models: 'with' and 'without' host alternation. From the Bayes factors obtained, the probabilities to re-estimate the true model are calculated for each life cycle (*i.e.* the model that is indeed used for the randomly drawn simulation). Additionally, we perform Principal Component Analyses (PCA) from the values of summary statistics obtained from the two life cycles. We display envelopes containing 95% of the simulations. A leave-one-out cross-validation for the parameter estimation is performed on 200 randomly drawn simulations from the 50,547 simulations under the life cycle 'with' host alternation.

Sampling schemes

In the first theoretical case presented in the results, we calculate summary statistics based on the maximum

information available. We name it the 'full' simulation design, as we consider all time samples, that is populations sampled once a year for 36 years on all host compartments (S and R for the model choice; S, R and A for the parameter estimation).

To assess the impact of the sampling scheme on inference accuracy, we consider two types of rarefaction affecting the time series and the range of compartments considered. Concerning the time series, in addition to the full time series, we consider (1) a thinning that keeps samplings every five years for a total of eight time samples and (2) only the first and last time samples. For the latter temporal rarefaction, it is not possible to compare the two types of summary statistic since only two generations are considered. Concerning rarefaction of the sampled compartment (S, R or A) we apply different rarefaction types depending on the cross-validation procedure.

The model choice procedure is based on: (1) Populations sampled on both S and R; (2) Populations sampled on S only. Populations sampled on A are not used for the model choice because this compartment only exists for the life cycle 'with' host alternation. The parameter estimation procedure focus on the life cycle 'with' host alternation and is based on: (1) Populations on S, R, and A; (2) Populations on S and R; (3) Populations on S and A; (4) Populations on S only.

2.4 Case study: overcoming of poplar rust resistance RMlp7

Here we apply our ABC inference framework to a documented case of resistance overcoming by a diploid plant pathogen responsible for the poplar rust disease, *M. larici-populina*. This pathogenic fungus is a host-alternating species. Its life cycle consists of an annual sexual reproduction on larch needles in early Spring, followed by clonal multiplications on poplar leaves from spring to autumn (Duplessis et al., 2021; Louet et al., 2023). Its sexual reproduction is obligatory in temperate climates because of the annual fall of poplar leaves (Xhaard, 2011). To control poplar rust disease, several resistant poplars have been selected and planted widely in Western Europe over years. However, the intensive and monocultural plantations combined with the host species being perennial (Gérard et al., 2006) makes poplars particularly vulnerable to the adaptation of the pathogen. This leads to regular events of resistance overcoming, so that all known resistance types have

now been overcome (Pinon and Frey, 2005; Louet et al., 2023). The most damaging resistance overcoming by *M. larici-populina* occurred in 1994 with the adaptation of the pathogen to the resistance RMlp7 carried at that time by a vast majority of cultivated poplar trees. The adaptation of the pathogen to the resistance RMlp7 resulted from the association of two alterations (a nonsynonymous mutation and a complete deletion) at the candidate locus *AvrMlp7*, from standing genetic variation (Louet et al., 2023). This led to a rapid invasion, in less than four years, of adapted pathogens across Western Europe, including France (Barrès et al., 2008; Xhaard et al., 2011; Persoons et al., 2017), causing drastic epidemics (Pinon et al., 1998; Pinon and Frey, 2005). This rapid adaptation event strongly shaped the resulting genetic structure of *M. larici-populina* populations (Xhaard et al., 2011; Persoons et al., 2017). This pathogen is airborne, which results in relatively high migration rates compared to telial organisms (Saubin et al., 2023a). This aerial dispersal also leads to high mortality during the annual migration on larch needles. Chosen prior distributions of parameters (Table 1) are consistent with our knowledge of the studied organism.

Poplar rust individuals were sampled before, during, and after this resistance overcoming event and correspond to the samples analysed in the population genetic study. The dataset used for our ABC analysis was previously studied (Louet et al., 2023), and was supplemented by including more recent samples to investigate temporal variation of population genetic indices over a longer period (Table S4). Genotyping of these additional samples was performed as described in Louet et al. (2023) and Persoons et al. (2017). Each sampled individual was genotyped with 20 microsatellite markers, and population genetic indices are calculated from the model developed above. Two versions of this temporal sampling are used. The first data set is composed of poplar rust samples from the same geographic location (Amance, France), collected from susceptible poplars (S) at the end of the epidemic season and from larch (A) after the sexual reproduction. Twenty-eight populations were sampled between 1989 and 2021 (22 on S, and 6 on A). Sampling size ranges from 5 to 58 individuals (samples with less than 5 individuals were removed from the analyses). The second data set is composed of poplar rust samples from a broader geographic region (Grand Est region, France), collected from susceptible and resistant poplars (S and R) at the end of the epidemic season, and from larch (A) after the sexual reproduction. This second data set includes all individuals from the first data set.

Thirty-six populations were sampled between 1988 and 2021 (26 on S, 4 on R and 6 on A). The sampling size ranges from 5 to 79 individuals (samples with less than 5 individuals were removed from the analyses). The empirical data used for the ABC inference are summarised in Table S4 and the total number of summary statistics in Table S2.

For this case study, cross-validation procedures are performed as described above, based on simulated data with the same sampling schemes as the two biological data sets. For the two data sets, the accuracy of the model choice and parameter estimation are compared depending on the choice of summary statistics (Complete summary statistics, or Wrap-up summary statistics). Model choice and parameter estimations are then performed as described above on the empirical data sets, and the posterior distribution is obtained for each model parameter. For each sampling scheme, we assess the relative importance of summary statistics to the inference of each parameter using the semi-automatic ABC method of Fearnhead and Prangle (2012) implemented in the `R` package `abctools` (Nunes and Prangle, 2015).

For the model choice, we check the goodness-of-fit by computing the $P-value$ to test the fit of the empirical data to each model. For each model, the null hypothesis states that the tested model offers a good fit and is rejected if $P-value < 0.01$.

3 Results

3.1 Model choice and parameter estimations for the 'full' simulation design

We first evaluate the accuracy of our ABC inference under the ideal case with the maximum data available: populations are sampled every generation from all host compartments.

The cross-validation procedure for the model choice highlights a strong accuracy of model choice when Wrap-up summary statistics are used (Table 3). Conversely, the model identifiability is weaker for Complete summary statistics. However, whatever the summary statistics used, there is a strong overlap in the outcome of simulations realised under the two life cycles (Figure 2). Hence only small areas of parameter values allow to truly discriminate between the two models.

The cross-validation procedure for parameter estimation highlights a strong discrepancy among parameters. Some parameters are very well estimated, including $propR$, K, f_{avr} and r, ranked in a decreasing order of identifiability (Table 4, Figure 3). τ is less accurately estimated, but with increasing confidence as its true value increases (Figure 3). Last, the worst parameter to estimate is mig, irrespective of the simulated value (Figure 3).

3.2 Model choice and parameter estimations when rarefaction is applied

For both cross-validations, sampling every five years gives similar results than considering the full time series of samplings. Reducing the range of sampled compartments also provides fairly good estimates for the model choice and parameter estimations. This result holds even if the sampling focuses on S only. Overall, the ABC accuracy for the model choice is higher for Wrap-up summary statistics than for Complete summary statistics (Table 5). The Wrap-up summary statistics performed equally well, or slightly better, when rarefaction is applied. It is noteworthy that the total number of Wrap-up summary statistics is not affected by the rarefaction, unlike for Complete summary statistics (Table S1). For parameter estimation, Wrap-up summary statistics generally perform better than for Complete summary statistics (Table 6). However, the differences in accuracy in the parameter estimations between Complete and Wrap-up summary statistics decreases with the rarefaction. Last, the accuracy of the model choice drops drastically if only the first and last time points are considered (Table 5) and the parameter estimations are less accurate, especially for r and τ. However $propR$, f_{avr} and K are still well estimated, regardless of the rarefaction (Table 6).

3.3 Case study: Approximate Bayesian Computations applied to the poplar rust resistance overcoming

In this section, we apply our ABC framework to the two empirical data sets describing the rapid evolution of the genetic structure of a pathogen population following resistance overcoming.

3.3.1 Cross-validations applied to the sampling schemes of empirical data sets

We first perform the cross-validation procedures with sampling schemes corresponding to the two data sets (Amance location and Grand Est region). For both data sets, the cross-validation of model choice shows weak identifiability of life cycles (Table 7). The identifiability of the life cycle 'with' host alternation is slightly better than 'without' host alternation but still limited. The model is more identifiable with the sampling scheme of the data set from the Grand Est region, which represents more time points and both S and R compartments, contrary to the data set from Amance location sampled on S only. As for the 'full' simulation design, the accuracy of parameter estimations is particularly high for parameters $propR$ and K, regardless of the data set (Table 8). The estimation of parameters f_{avr} is more mitigated but still quite good (correlation coefficient > 0.75 for the two data sets). Conversely, mig, r and τ are less accurately estimated. Unlike the 'full' simulation design, there are no clear differences of accuracy in the inference of parameters using Wrap-up summary statistics or Complete summary statistics.

The population genetic index that contributes most to the inference is TF_{ST}, the population differentiation between the initial population (after burn-in) and the sampled population (Tables S3, S5). Considering the Complete summary statistics, TF_{ST} with sampling at generation 8 exhibits the major contribution to the inference of parameters $propR$, r, K and τ for both sampling schemes (Table S3). Considering the Wrap-up summary statistics, this information is encapsulated in $VarTF_{ST}$. Note that for the Grand Est sampling scheme, $VarTF_{ST-R}$ exhibits the major contribution to most inferred parameters, which highlights the importance of sampling on R (Table S5).

3.3.2 Model choice

For the two empirical data sets, we infer the life cycle 'with' host alternation with probability 100% or 97% from the Complete summary statistics and 100% or 99% from the Wrap-up summary statistics, depending on the empirical data set (Table 9). In all cases however, the results of the goodness-of-fit tests do not allow to significantly reject either of the life cycles, irrespective of the data set considered and the summary statistics used. Therefore, we do not have enough information to properly infer the pathogen life cycle. This

is consistent with the coordinates of the data sets on the PCA analyses (Figure 4), which are located in the overlap area of the two models. This explains the weak identifiability of the life cycle and the impossibility of significantly rejecting either life cycle. Therefore, in our case study we have to rely on biological insights to determine the life cycle.

3.3.3 Parameter estimation

Knowing that the poplar rust pathogen alternates on larch to perform its sexual reproduction, we compute the parameter estimations under the model 'with' host alternation.

In the case study, Complete summary statistics perform better than Wrap-up summary statistics for the analysis of both empirical data sets. For most parameters, especially those that are well identified, the difference between the posterior density versus the prior density is more pronounced (Figures 5, S1). Therefore, in the following, we focus on the parameter estimation from Complete summary statistics. As expected from the theoretical identifiability, narrow posterior distributions allow a confident inference of parameters $propR$ and K. The power of inference of parameters f_{avr}, r and τ is more limited, and the posterior distribution for parameter mig is not more informative compared to its prior distribution. From the mode of each posterior distribution (Table 10), we infer a high proportion of resistant poplars at the time of resistance overcoming: $propR = 0.93$ and $propR = 0.85$ for Amance and the Grand Est region data sets respectively. The inferred population sizes are in the order of three thousands (mode values $K = 3,751$ and $3,404$ for Amance and the Grand Est region, respectively). We infer an initial proportion of virulent alleles f_{avr} in the pathogen population of 13% in Amance and in the Grand Est region, but with a large credible interval, between 3% and 20%. The annual mortality rate τ during the annual migration preceding sexual reproduction is inferred at 0.56 and 0.75 in Amance and in the Grand region, respectively. The growth rate r is estimated at 1.63 and 1.96 in Amance and in the Grand region, respectively. The migration rate mig is badly estimated and would range from 0.01 to 0.10.

4 Discussion

Time samples unravel rapid adaptation

In this paper, we develop an original approach to infer demographic scenarios and parameter values in the case of rapid adaptation. We base our inference framework on the use of time series data to grasp changes in population structure over time. We employ a forward population genetic model to simulate the summary statistics used for ABC inference. As such, our study is in line with the most recent developments in population genetic inference that use time series data to unravel the demographic changes that accompany selection (Foll et al., 2015; Johri et al., 2022; Kreiner and Booker, 2022; Pavinato et al., 2022). Theoretical investigations demonstrate high accuracy in both model and parameter estimations. Last, our inferential framework is successfully applied to a case study of rapid adaptation in a plant pathogen following a resistance overcoming event with consistent estimates of demographic and ecological oriented parameters.

Too many summary statistics lead to model overfitting

The main originality of our method is to take into account the time series, compared to the high amount of studies that focus on a single time point. To do so, we adapt the classical ABC protocols of using summary statistics. We compare two methods, either keeping the complete sequence of statistics over all time samples or wrapping up the information into the mean, variance, minimum and maximum values for each summary statistic. We show in the theoretical cross-validation procedures that Wrap-up summary statistics outperform Complete summary statistics, except when we apply data rarefaction. This result may appear counter-intuitive. However too much and redundant information can lead to a decrease in statistical power. Since many statistics can be used in ABC, there is a certain curse of dimensionality, early on identified. This generates overfitting that may affect the accuracy of model and parameter estimation. This is especially acute when several summary statistics correlate with the same model parameters, inflating the heterogeneous accuracy inference of some parameters compared to others. A suggested solution is to perform first a dimension reduction of the statistics space (Wegmann et al., 2009) before performing the ABC estimations. In our case, we have few genetic indices and few parameters, so we do not perform such reduction of dimensionality

beforehand. We want to keep the full time series to avoid loss of temporal information, but in doing so we inflate the number of summary statistics which is likely to result in temporal auto-correlation causing overfitting. In support of this hypothesis, the discrepancy in parameter identifiability between Complete and Wrap-up statistics tends to decrease when we apply a thinning in the time series consisting in keeping only one fifth of the temporal samples. A way to keep the maximum information while avoiding the redundancy that causes overfitting would be to summarise the time series in a different manner. For example, we may fit a function describing the temporal changes and use regression coefficients as summary statistics. However, this method requires that all dynamics be correctly fitted by a similar function (with the same number of regression coefficients). In this attempt, we try to fit quadratic polynomes, but the shape of the temporal variation varied too much among population genetic indices to provide conclusive summary statistics (data not shown). Moreover, such fit cannot take into account stochastic variation due to drift. Another way to improve the inferential framework would be to avoid the use of population genetic indices as summary statistics and base the inference method directly on the evolution of genotypic frequencies, or a less condensed type of information such as site frequency spectra (SFS) across time samples. However this method may be more computationally intensive (in particular for storing the results over a long time series), and the SFS cannot be computed from microsatellite markers. Moreover, the resulting increase in the number of summary statistics could also lead to high dimensionality of the data.

Case study estimates match the biology of the poplar rust system

By applying this method on empirical data sets, we infer accurately three parameters: $propR$, K and f_{avr}. The estimated values make sense with regard to our knowledge of this event of resistance overcoming. We infer a very large proportion of resistant poplars (85% to 93%), which is slightly higher but still consistent with our knowledge of poplar plantations before the RMlp7 resistance overcoming. Indeed, the resistant poplar cultivar 'Beaupré' that bears resistance RMlp7 was widely planted at the time of resistance overcoming and represented up to 80% of poplar cultivar sales in 1996 (data from the French Ministry of Agriculture, Fabre et al., 2021). This very high proportion of resistant poplars in the landscape exerted a strong selection

pressure and accounts for the changes in the genetic structure of a pathogen population over time (Persoons et al., 2017). We infer a population size of around three thousand individuals. This population size does not represent the number of individuals actually multiplying in the population but the effective population size, *i.e.* the number of individuals that effectively contribute to the observed genetic variability. This order of magnitude is consistent with a previous estimate based on a coalescent analysis (Persoons, 2015). We estimate an initial proportion of virulent alleles of 13%. This value is consistent with the genetic characterisation at the selected locus in *M. larici populina* (Louet et al., 2023). Louet et al. (2023) highlighted that the alleles conferring virulence pre-existed in the pathogen population long before the resistance overcoming, at a frequency of 21%, on average, between 1989 and 1993. The proportion of virulent alleles in the population can strongly fluctuate during the years preceding resistance overcoming (Saubin et al., 2021). Such fluctuations are indeed observed in the data from Louet et al. (2023), with a standard deviation in the proportion of virulent alleles of 0.12 between 1989 and 1993. Our estimation is therefore consistent with the empirical data considering such large fluctuations from year to year. This level of standing genetic variation is not negligible in view of the adaptive potential it brings to the pathogen population.

The fact that some parameters are not well estimated is not caused by a limited amount of data, but rather by a strong assumption in our modelling framework. Indeed, we assume three compartments only in the landscape and neglect the more complex spatial structures that can be encountered in agricultural landscapes. In such a non-spatial system, it may be especially difficult to disentangle the genetic effects of growth, mortality and migration rates. We believe that increasing the spatial complexity of the model would help disentangle these three parameters. However, this would also highly increase the dimensions of summary statistics, which can lead to a decrease in the statistical accuracy of the ABC.

Despite the precautions taken when evaluating the theoretical accuracy of our method, we observe that the accuracy of our ABC estimation differs between the theoretical and empirical results. We obtain similar theoretical results with Wrap-up summary statistics and with Complete summary statistics. However, Complete summary statistics lead to more accurate parameter inferences from empirical data sets. We believe that this is due to increased variability in population genetic indices in the empirical data compared to the

simulated data. Many sources of biological stochasticity are not accounted for in our modelling assumptions. This stochasticity results in stronger local extremes, which are captured by the minimum and maximum values in Wrap-up summary statistics. Thus, taking into account all the available information of the empirical time series through Complete summary statistics can reduce the impact of this additional stochasticity, and improve the parameter estimations. This discrepancy between our application to empirical data and theoretical findings from simulated data can also originate from strong modelling assumptions (like the non-spatial model) that do not perfectly reflect the biological system.

Methodological guideline

We recommend using a sampling scheme with regular time points to capture sufficient information in the genetic evolution of populations. Contrary to the classical practices, we show that for this framework it is more efficient to sample more populations in time but on a single host than to focus on few temporal samples but on several hosts. The sampling scheme with the first and the last time points, even if less informative than a more regular sampling, still allows to correctly infer some of the model parameters. In particular, the initial frequency of virulent alleles in the population is very well inferred with this sampling scheme because the first population in time is the most important to identify the initial genetic composition of the population. In addition to the methodological considerations of ABC analyses, it is essential to build a model adapted to the organism studied. For example, an organism with a high clonality rate may lead to special methodological considerations, with a much longer burn-in period required before a steady state is reached (Reichel et al., 2016; Hartfield, 2021).

Future work is still required to establish a robust and general framework for demography inference of species with rapid adaptive evolution. It would be interesting to test this framework on another species with a life cycle without host alternation. This analysis would serve as a second proof of concept without the need to modify the underlying simulation model. In addition, further developments could benefit from modelling spatially structured populations more realistically. This would lead assuredly to a more generic model and would allow for a better estimation of the migration rate. Such an addition could however increase

the simulation time, which is a major issue in ABC analyses where the number of simulations can be decisive for increasing inferential power.

Future directions: application for genome scans of rapid adaptation

Our analyses demonstrate that neutral genetic data from microsatellite markers contain sufficient information to infer parameters of resistance overcoming from time samples. This validates our approach consisting in studying rapid adaptation in two steps: first inferring demographic parameters using neutral loci only, then apply the demographic scenario to build accurate genome scan analyses. The next step thus involve considering population genomic data. The integration of genomic data would allow the calculation of a wider range of summary statistics, which could increase the accuracy of inferences. Fitting demographic models is a prerequisite for detecting selection, but it can be difficult to do in practice and is often inaccurate (Hoban et al., 2016). The addition of genomic data to the described framework would enable to focus on areas under selection, detect sweeps (Messer and Petrov, 2013; Foll et al., 2015) and calculate their age. Determining the evolution of neutral loci through such a selection event may allow, by comparison, to identify selected loci implied in rapid adaptation.

5 Acknowledgements

We warmly thank Bénédicte Fabre and Jérémy Pétrowski for collecting and phenotying the additional *M. larici-populina* populations, Emma Chavan and Lola Mottet for their help with the microsatellite analyses, and Maria Orive for constructive comments on a previous version of the manuscript. This work was supported by grants from the French National Research Agency (ANR-18-CE32-0001, Clonix2D project; ANR-11-LABX-0002-01, Cluster of Excellence Arbre) and from the Metaprogram SuMCrop of the National Research Institute for Agriculture, Food and the Environment (INRAE, Opiniâtres project). Méline Saubin was supported by a PhD fellowship from INRAE and the French National Research Agency (ANR-18-CE32-0001, Clonix2D project). Méline Saubin obtained an international mobility grant BayFrance as part of a Franco-Bavarian cooperation project, to work for one month in Aurélien Tellier's lab (Grant Number

FK21_2020).

References

Agapow, P. M. and Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, 1(1-2):101–102.

Agrios, G. N. (2005). *Plant pathology*. Elsevier edition.

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., and Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology and Evolution*, 19(10):536–544.

Arnaud-Haond, S., Duarte, C. M., Alberto, F., and Serrão, E. A. (2007). Standardizing methods to address clonality in population studies. *Molecular Ecology*, 16:5115–5139.

Barrès, B., Halkett, F., Dutech, C., Andrieux, A., Pinon, J., and Frey, P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: Evidence for isolation by distance in Europe and recent founder effects overseas. *Infection, Genetics and Evolution*, 8(5):577–587.

Bazin, E., Dawson, K. J., and Beaumont, M. A. (2010). Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, 185(2):587–602.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., and Petrov, D. A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genetics*, 10(11).

Bessho, K. and Otto, S. P. (2022). Fixation and effective size in a haploid–diploid population with asexual reproduction. *Theoretical Population Biology*, 143:30–45.

Blum, M. G. B. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73.

Bonneaud, C. and Longdon, B. (2020). Using evolutionary theory to understand the fate of novel infectious pathogens. *Science and Society*, 21(e51374).

Brown, J. K. M. and Tellier, A. (2011). Plant-parasite coevolution: Bridging the gap between genetics and ecology. *Annual Review of Phytopathology*, 49(1):345–367.

Buffalo, V. and Coop, G. (2019). *The linked selection signature of rapid adaptation in temporal genomic data*, volume 213.

Buffalo, V. and Coop, G. (2020). Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences of the United States of America*, 117(34):20672–20680.

Burdon, J. J., Zhan, J., Barrett, L. G., Papaix, J., and Thrall, P. H. (2016). Addressing the challenges of pathogen evolution on the world's arable crops. *Phytopathology*, 106(10):1117–1127.

Ceplitis, A. (2003). Coalescence times and the Meselson effect in asexual eukaryotes. *Genetical Research*, 82(3):183–190.

Collin, F.-D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J.-M., and Estoup, A. (2021). Extending approximate bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using diyabc random forest. *Molecular Ecology Resources*, 21(8):2598–2613.

Cornuet, J. M., Ravigné, V., and Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, 11(401).

Csilléry, K., François, O., and Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479.

Day, T. and Gandon, S. (2007). Applying population-genetic models in theoretical evolutionary epidemiology. *Ecology Letters*, 10:876–888.

De Mita, S. and Siol, M. (2012). EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, 13(1):27.

De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., and Vigouroux, Y. (2013). Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, 22(5):1383–1399.

Dehasque, M., Ávila-Arcos, M. C., Díez-del Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E. D., Malaspinas, A. S., Marques-Bonet, T., Martin, M. D., Murray, G. G. R., Papadopulos, A. S. T., Therkildsen, N. O., Wegmann, D., Dalén, L., and Foote, A. D. (2020). Inference of natural selection from ancient DNA. *Evolution Letters*, 4(2):94–108.

Dobzhansky, H. (1943). Genetics of natural populations IX. Temporal changes in the composition of populations of *Drosophila pseudoobscura*. *Genetics*, 28:162–186.

Dobzhansky, T. and Pavlovsky, O. (1971). Experimentally created incipient species of *drosophila*. *Nature*, 230(5292):289–292.

Duplessis, S., Lorrain, C., Petre, B., Figueroa, M., Dodds, P. N., and Aime, M. C. (2021). Host adaptation and virulence in heteroecious rust fungi. *Annual Review of Phytopathology*, 59:403–422.

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution.

Estoup, A. and Guillemaud, T. (2010). Reconstructing routes of invasion using genetic data: Why, how and so what? *Molecular Ecology*, 19(19):4113–4130.

Fabre, B., Bastien, C., Husson, C., Marçais, B., Frey, P., and Halkett, F. (2021). Chapitre 27. Un effet papillon dans les peupleraies françaises : les répercussions d'un contournement de résistance sur les méthodes de sélection variétale. In Lannou, C., Roby, D., Ravigné, V., Hannachi, M., and Moury, B., editors, *L'immunité des plantes*, pages 329–339. Versailles, quae edition.

Fagundes, N. J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17614–17619.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(3):419–474.

Feder, A. F., Pennings, P. S., and Petrov, D. A. (2021). The clarifying role of time series data in the population genetics of HIV. *PLoS Genetics*, 17(1):1–10.

Fisher, R. A. and Ford, E. B. (1947). The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity*, 1(2):143–174.

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annual Review of Phytopathology*, 9(1):275–296.

Foll, M., Shim, H., and Jensen, J. D. (2015). WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1):87–98.

Gérard, P. R., Husson, C., Pinon, J., and Frey, P. (2006). Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. *Phytopathology*, 96(9):1027–1036.

Hacquard, S., Petre, B., Frey, P., Hecker, A., Rouhier, N., and Duplessis, S. (2011). The poplar-poplar rust interaction: Insights from genomics and transcriptomics. *Journal of Pathogens*, 2011:1–11.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hartfield, M. (2021). Approximating the Coalescent Under Facultative Sex. *The Journal of heredity*, 112(1):145–154.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., and Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, 188(4):379–397.

Jin, Y., Szabo, L. J., and Carson, M. (2010). Century-old mystery of *Puccinia striiformis* life history solved with the identification of berberis as an alternate host. *Phytopathology*, 100(5):432–435.

Johnson, R. (1984). A critical analysis of durable resistance. *Annual Review of Phytopathology*, 22.

Johri, P., Aquadro, C. F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., Keightley, P. D., Lynch, M., McVean, G., Payseur, B. A., Pfeifer, S. P., Stephan, W., and Jensen, J. D. (2022). Recommendations for improving statistical inference in population genomics. *PLoS Biology*, 20(5):1–23.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kettlewell, H. B. D. (1958). A survey of the frequencies of *Biston betularia* (L.) (Lep.) and its melanic forms in great britain. *Heredity (Edinb.)*, 12:51–72.

Kettlewell, H. B. D. (1961). The phenomenon of industrial melanism in Lepidoptera. *Annual Review of Entomology*, 6(1):245–262.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Kreiner, J. M. and Booker, T. R. (2022). Disentangling the genetic consequences of demographic change. *Molecular Ecology*, 32:278–280.

Laval, G., Patin, E., Boutillier, P., and Quintant-Murci, L. (2019). A genome-wide Approximate Bayesian Computation approach suggests only limited numbers of soft sweeps in humans over the last 100,000 years. *bioRxiv*, pages 1–53.

Louet, C., Saubin, M., Andrieux, A., Persoons, A., Gorse, M., Pétrowski, J., Fabre, B., De Mita, S., Duplessis, S., Frey, P., and Halkett, F. (2023). A point mutation and large deletion at the candidate avirulence locus *AvrMlp7* in the poplar rust fungus correlate with poplar RMlp7 resistance breakdown. *Molecular Ecology*, (February):1–12.

Luqman, H., Widmer, A., Fior, S., and Wegmann, D. (2021). Identifying loci under selection via explicit demographic models. *Molecular Ecology*, 21(8):2719–2737.

McDonald, B. A. and Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annual Review of Phytopathology*, 40:349–379.

Messer, P. W. and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, 28(11):659–669.

Mueller, L. D., Barr, L. G., and Ayala, F. J. (1985a). Natural selection vs. random drift: Evidence from temporal variation in allele frequencies in nature. *Genetics*, 111(3):517–554.

Mueller, L. D., Wilcox, B. A., Ehrlich, P. R., Heckel, D. G., and Murphy, D. D. (1985b). A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha*. *Genetics*, 110(3):495–511.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3):583–590.

Nei, M. and Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics*, 98(3):625–640.

Nielsen, R. and Signorovitch, J. (2003). Correcting for ascertainment biases when analyzing SNP data: Applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, 63(3):245–255.

Nunes, M. A. and Prangle, D. (2015). abctools: An R package for tuning Approximate Bayesian Computation analyses. Forthcoming.

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365:185–205.

Orive, M. E. (1993). Effective population size in organisms with complex life-histories.

Parat, F., Schwertfirm, G., Rudolph, U., Miedaner, T., Korzun, V., Bauer, E., Schön, C. C., and Tellier, A. (2016). Geography and end use drive the diversification of worldwide winter rye populations. *Molecular Ecology*, 25(2):500–514.

Parsons, T. L., Lambert, A., Day, T., and Gandon, S. (2018). Pathogen evolution in finite populations : slow and steady spreads the best. *Journal of royal society*, 15.

Pavinato, V. A. C., De Mita, S., Marin, J. M., and De Navascués, M. (2022). Joint inference of adaptive and demographic history from temporal population genomic data. *Peer Community Journal*, 2(e78):1–20.

Persoons, A. (2015). *Les contournements de résistance par Melampsora larici-populina, l'agent de la rouille du peuplier : impact démographique et déterminisme génétique*. PhD thesis, Université de Lorraine.

Persoons, A., Hayden, K. J., Fabre, B., Frey, P., De Mita, S., Tellier, A., and Halkett, F. (2017). The escalatory Red Queen: Population extinction and replacement following arms race dynamics in poplar rust. *Molecular Ecology*, 26(7):1902–1918.

Pink, D. and Puddephat, I. (1999). Deployment of disease resistance genes by plant transformation - a 'mix and match' approach. *Trends in Plant Science*, 4(2):71–75.

Pinon, J. and Frey, P. (2005). Interactions between poplar clone and *Melampsora* populations and their implications for breeding for durable resistance. *Rust Diseases of Willow and Poplar*, (July):138–154.

Pinon, J., Frey, P., Husson, C., and Schipfer, A. (1998). Poplar rust (*Melampsora larici-populina*): the development of E4 pathotypes in France since 1994. *First IUFRO Rusts of Forest Trees*, 712:57–64.

Pollak, E. (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics*, 104(3):531–548.

Prout, T. (1954). Genetic drift in irradiated experimental populations of *Drosophila melanogaster*. *Genetics*, 39(4):529–545.

Reichel, K., Masson, J. P., Malrieu, F., Arnaud-Haond, S., and Stoeckel, S. (2016). Rare sex or out of reach equilibrium? The dynamics of FIS in partially clonal organisms. *BMC Genetics*.

Rimbaud, L., Fabre, F., Papaïx, J., Moury, B., Lannou, C., Barrett, L. G., and Thrall, P. H. (2021). Models of plant resistance deployment. *Annual Review of Phytopathology*, 59(1).

Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390.

Saubin, M., Coville, J., Xhaard, C., Frey, P., Soubeyrand, S., Halkett, F., and Fabre, F. (2023a). Inferring invasion determinants with mechanistic models and multitype samples. *bioRxiv*.

Saubin, M., De Mita, S., Zhu, X., Sudret, B., and Halkett, F. (2021). Impact of ploidy and pathogen life cycle on resistance durability. *Peer Community Journal*, 1(e8):1–31.

Saubin, M., Louet, C., Bousset, L., Fabre, F., Frey, P., Fudal, I., Grognard, F., Hamelin, F., Mailleret, L., Stoeckel, S., Touzeau, S., Petre, B., and Halkett, F. (2023b). Improving sustainable crop protection using population genetics concepts. *Molecular Ecology*, 00:1–11.

Saubin, M., Stoeckel, S., Tellier, A., and Halkett, F. (2022). Interplay between demography and selection: A forward model illuminates the temporal genetic signatures of rapid adaptation in plant pathogens. *bioRxiv, en révision pour Heredity*.

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., Mcroberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature ecology & evolution*, 3(3):430–439.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–1070.

Stukenbrock, E. H. and McDonald, B. A. (2008). The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology*, 46:75–100.

Talhinhas, P., Batista, D., Diniz, I., Vieira, A., Silva, D. N., Loureiro, A., Tavares, S., Pereira, A. P., Azinheira, H. G., Guerra-Guimarães, L., Várzea, V., and Silva, M. (2017). The coffee leaf rust pathogen *Hemileia vastatrix*: one and a half centuries around the tropics. *Molecular Plant Pathology*, 18(8):1039–1051.

Tellier, A. and Lemaire, C. (2014). Coalescence 2.0: A multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11):2637–2652.

Tobin, P. C. (2015). Ecological consequences of pathogen and insect invasions. *Forest pathology and entomology*, 1:25–32.

van Rossum, G. (1995). Python tutorial, Technical Report CS-R9526. *CWI*.

Vitalis, R., Gautier, M., Dawson, K. J., and Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, 196(3):799–817.

Wallace, B. (1956). Studies on irradiated populations of *Drosophila melanogaster*. *Journal of Genetics*, 54(2):280–293.

Wang, J. and Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1):429–446.

Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218.

Whitlock, M. C. and Barton, N. H. (1997). The effective size of subdivided population. *Genetics*, 146(1):427–441.

Wright, S. (1949). Adaptation and selection. In G. L. Jepson, G. G. S. and E. Mayr, E., editors, *Genetics, Paleontology, and Evolution*, chapter Adaptation, pages 365–389. Princeton Univ. Press, Princeton, NJ.

Wright, S. (1978). *Evolution and the genetics of populations - Variability within and among natural populations*. The University of Chicago Press.

Xhaard, C. (2011). *Influence des processus démographiques sur la structure et les caractéristiques génétiques des champignons phytopathogènes, cas de l'agent de la rouille du peuplier Melampsora larici-populina*. PhD thesis.

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barrès, B., Frey, P., and Halkett, F. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. *Molecular Ecology*, 20(13):2739–2755.

Zhan, J., Thrall, P. H., Papaïx, J., Xie, L., and Burdon, J. J. (2015). Playing on a pathogen's weakness: Using evolution to guide sustainable plant disease control strategies. *Annual Review of Phytopathology*, 53(1):19–43.

Živković, D., John, S., Verin, M., Stephan, W., and Tellier, A. (2019). Neutral genomic signatures of host-parasite coevolution. *BMC Evolutionary Biology*, 19(1):1–11.

Data accessibility

`R` scripts for statistical analyses and data for the biological application as well as an executable file to run the population genetic simulations are available on a public GitLab repository: https://gitlab.com/saubin.meline/demogenetic-abc.

Author contributions

Méline Saubin, Aurélien Tellier and Fabien Halkett conceived and designed the study. Axelle Andrieux performed the additionnal genotyping. Méline Saubin and Solenn Stoeckel produced the code and ran the simulations. Méline Saubin, Aurélien Tellier and Fabien Halkett analysed the data and prepared the manuscript. All authors revised and approved the manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Tables and Figures

Table 1: Input parameters and their range of variations for the random simulation design run for 400 generations.

Parameter	Description	Distribution	Interval
f_{avr}	Initial frequency of the virulent allele in the pathogen population	Uniform	[0.01; 0.2]
$Cycle$	Life cycle of the pathogen	Bernoulli	Without or 'with' host alternation, probability 0.5
mig	Migration rate between R and S	Uniform	[0.01; 0.1]
r	Growth rate of the pathogen population	Uniform	[1.1; 2]
τ	Mortality rate during the annual migration	Uniform	[0.5; 1[
K	Cumulative carrying capacity of susceptible and resistant hosts	Uniform	[1,000; 20,000]
$propR$	Proportion of resistant hosts in the landscape	Uniform	]0.01; 0.99[

Table 2: Description of population genetic indices used as summary statistics in the ABC analyses.

Index	**Description**	**Reference**
P_{Vir}	Proportion of virulent individuals	
β_P	Genotypic diversity index	Arnaud-Haond et al., 2007
R	Proportion of unique genotypes	
S	Simpson index of genetic diversity	
SW	Shannon-Wiener index of genetic diversity	
MH_E	Mean expected heterozygosity overall loci	Nei, 1978
VH_E	Variance of the expected heterozygosity overall loci	
MLA	Mean number of alleles by locus	Nielsen and Signorovitch, 2003
$\bar{r}_D$	Linkage disequilibrium index	Agapow and Burt, 2001
F_{ST} $R-S$	Population differentiation between populations on R and S	Wright, 1949, 1978
TF_{ST}	Population differentiation between the initial population (after burn-in) and the sampled population	

Table 3: Accuracy of model selection for the 'full' simulation design depending on the type of summary statistics considered. The model choice procedure is based on leave-one-out cross-validations with a weighted multinomial logistic regression computed with tolerance parameter set at 0.05, for 500 replicates.

Summary statistics	Probability to find the true model	
	'Without' host alternation	'With' host alternation
Complete	0.80	0.89
Wrap-up	1.00	0.99

Table 4: Accuracy of parameter estimation for the 'full' simulation design depending on the summary statistics considered. Data represent correlation coefficients between simulated and estimated parameters. The parameters identifiability procedure is based on a leave-one-out cross-validation with the neural networks regression method and tolerance parameter set at 0.05, for 200 replicates.

Summary statistics	$propR$	mig	f_{avr}	r	K	τ
Complete	0.91	0.42	0.77	0.74	0.76	0.51
Wrap-up	0.92	0.62	0.82	0.79	0.87	0.76

Table 5: Accuracy of model selection depending on the summary statistics and the sampling rarefaction considered. The model choice procedure is based on leave-one-out cross-validations with a weighted multinomial logistic regression computed with tolerance parameter set at 0.05, for 500 replicates. Bold values represent the values already presented in Table 3.

Time samples	Compartment samples	Summary statistics	Probability to find the true model	
			Without alternation	With alternation
Every year	S and R	Complete	**0.80**	**0.89**
		Wrap-up	**1.00**	**0.99**
	S	Complete	0.59	0.77
		Wrap-up	0.94	0.93
Every five year	S and R	Complete	0.76	0.81
		Wrap-up	1.00	0.99
	S	Complete	0.61	0.66
		Wrap-up	0.97	0.96
First and last year	S and R	-	0.59	0.62
	S	-	0.53	0.60

Table 6: Accuracy of the parameter estimation depending on the summary statistics considered and the type of rarefaction, for simulated populations 'with' host alternation. Data represent correlation coefficients between simulated and estimated parameters. The parameters identifiability procedure is based on a leave-one-out cross-validation with the neural networks regression method and tolerance parameter set at 0.05, for 200 replicates. Bold values represent the values already presented in Table 4.

Time samples	Compartment samples	Summary statistics	$propR$	mig	f_{avr}	r	K	τ
Every year	S, R and A	Complete	**0.91**	**0.42**	**0.77**	**0.74**	**0.76**	**0.51**
		Wrap-up	**0.92**	**0.62**	**0.82**	**0.79**	**0.87**	**0.76**
	S and R	Complete	0.90	0.43	0.77	0.74	0.79	0.53
		Wrap-up	0.90	0.51	0.85	0.76	0.86	0.68
	S and A	Complete	0.90	0.31	0.74	0.72	0.77	0.61
		Wrap-up	0.93	0.46	0.76	0.77	0.87	0.73
	S	Complete	0.87	0.29	0.73	0.61	0.79	0.60
		Wrap-up	0.90	0.42	0.74	0.63	0.80	0.64
Every five year	S, R and A	Complete	0.91	0.40	0.76	0.76	0.71	0.64
		Wrap-up	0.79	0.42	0.75	0.58	0.73	0.72
	S and R	Complete	0.88	0.33	0.82	0.61	0.81	0.61
		Wrap-up	0.92	0.55	0.81	0.60	0.88	0.69
	S and A	Complete	0.89	0.19	0.79	0.67	0.83	0.68
		Wrap-up	0.81	0.17	0.76	0.57	0.75	0.68
	S	Complete	0.86	0.21	0.75	0.59	0.83	0.51
		Wrap-up	0.87	0.38	0.70	0.60	0.78	0.59
First and last year	S, R and A	-	0.74	0.41	0.84	0.40	0.77	0.59
	S and R	-	0.78	0.49	0.76	0.39	0.68	0.50
	S and A	-	0.80	0.25	0.78	0.51	0.80	0.70
	S	-	0.72	0.19	0.79	0.18	0.79	0.56

Table 7: Accuracy of model selection on simulated populations on S and R, depending on the summary statistics considered. The sampling schemes of simulation data sets match those of the two empirical data sets: Amance location and Grand Est region. The model choice procedure is based on leave-one-out cross-validations with a weighted multinomial logistic regression computed with tolerance parameter set at 0.05, for 500 replicates.

Sampling scheme	Summary statistics	Probability to find the true model	
		'Without' host alternation	'With' host alternation
Amance	Complete	0.65	0.77
	Wrap-up	0.70	0.79
Grand Est	Complete	0.75	0.82
	Wrap-up	0.79	0.87

Table 8: Accuracy of the parameter estimation on simulated populations sampled on S, R and A compartments for the life cycle 'with' host alternation, depending on the summary statistics considered. The sampling schemes of the simulation data sets match those of the two the empirical data sets: Amance location and Grand Est region. Numbers represent correlation coefficients between simulated and estimated parameters. The parameter identifiability procedure is based on a leave-one-out cross-validation with the neural networks regression method and tolerance parameter set at 0.05, for 200 replicates.

Sampling scheme	Summary statistics	$propR$	mig	f_{avr}	r	K	τ
Amance	Complete	0.91	0.36	0.80	0.70	0.81	0.63
	Wrap-up	0.88	0.40	0.76	0.72	0.85	0.61
Grand Est	Complete	0.90	0.39	0.76	0.68	0.82	0.64
	Wrap-up	0.86	0.33	0.78	0.68	0.81	0.59

Table 9: Posterior model probabilities and goodness of fit for two data sets: in Amance location and the Grand Est region, depending on the summary statistics. 'With' and 'Without' stand for life cycles 'with' and 'without' host alternation, respectively. $P - values < 0.01$ are considered significant.

Sampling scheme	Summary statistics	Posterior probability		Goodness of fit ($P - value$)	
		Without	With	Without	With
Amance	Complete	0.00	1.00	0.084	0.045
	Wrap-up	0.00	1.00	0.053	0.019
Grand Est	Complete	0.03	0.97	0.086	0.056
	Wrap-up	0.01	0.99	0.073	0.085

Table 10: Statistical summary of the inference of the parameters for the life cycle 'with' host alternation, depending on the data set, with Complete summary statistics.

Data set	$Parameter$	$q - 2.5\%$	$median$	$mean$	$mode$	$q - 97.5\%$
Amance	$propR$	0.57	0.89	0.86	**0.93**	0.97
	mig	0.01	0.06	0.06	**0.09**	0.10
	f_{avr}	0.03	0.13	0.12	**0.13**	0.20
	r	1.30	1.71	1.70	**1.63**	2.01
	K	1,642	4,447	5,566	**3,751**	14,271
	τ	0.51	0.69	0.70	**0.56**	0.92
Grand Est	$propR$	0.43	0.81	0.79	**0.85**	0.93
	mig	0.01	0.06	0.06	**0.04**	0.10
	f_{avr}	0.03	0.12	0.12	**0.13**	0.19
	r	1.30	1.70	1.69	**1.96**	1.99
	K	1,010	3,976	4,942	**3,404**	13,253
	τ	0.51	0.70	0.71	**0.75**	0.92

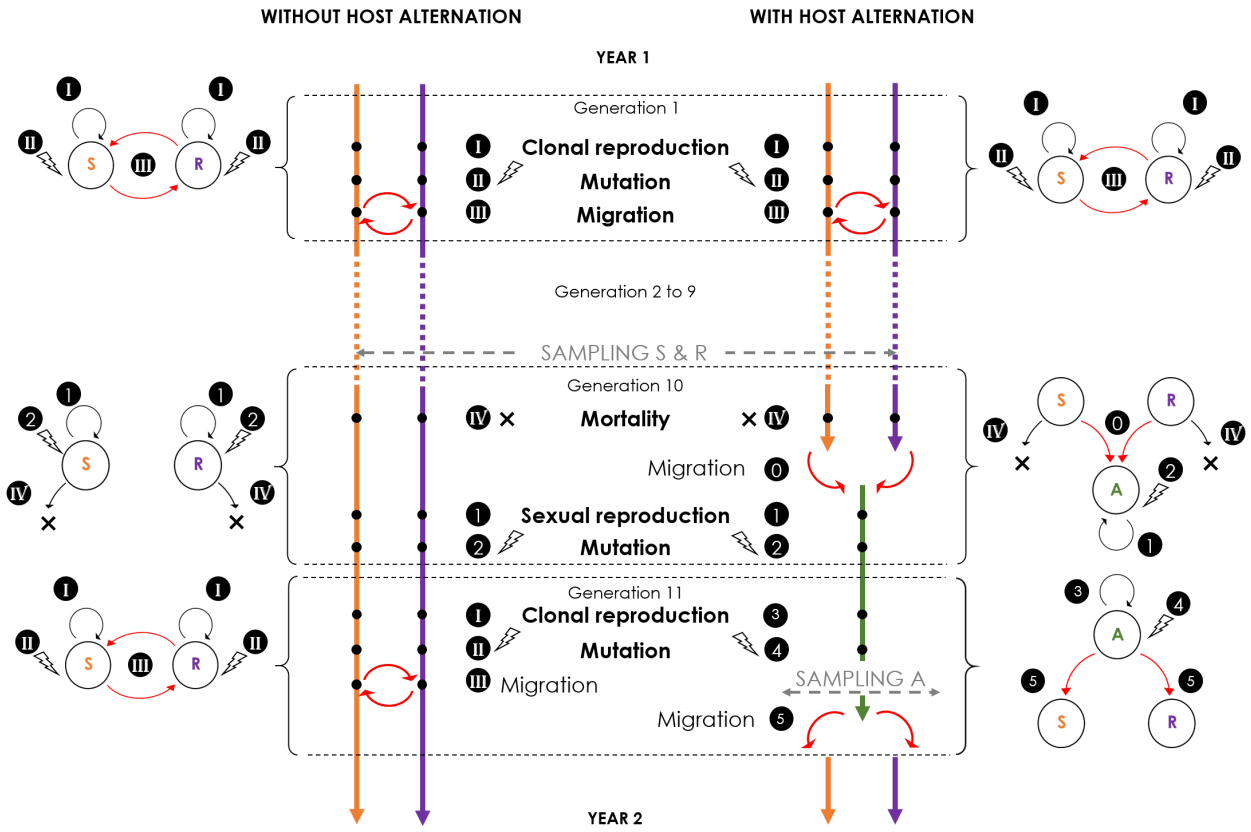
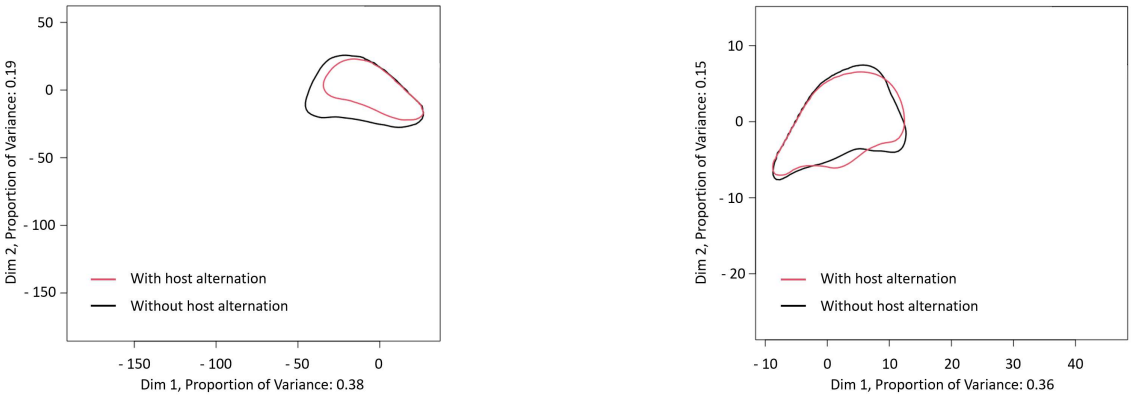


Figure 1: Modelling steps for each simulated year with the three S, R, and A host compartments. Each year is composed of 11 generations. During the clonal phase (generation 1 to 9), each generation is composed of three steps identical between both life cycles: (I) clonal reproduction; (II) migration of a proportion *mig* of each population between R and S; (III) mutation at all neutral markers with a mutation rate of 10^{-3}. At the end of the clonal phase, the pathogen overwinter as a dormant stage and is subjected to (IV) mortality of a proportion τ of each population. Then, the sexual phase (generation 10) differs depending on the life cycle: (0) represents the migration of all individuals from R and S towards A; (1) sexual reproduction; (2) mutation of all neutral markers with a mutation rate of 10^{-3}. This sexual phase is followed by a new clonal phase, which is identical 'without' host alternation to the first clonal phase and 'with' host alternation: (3) represents the clonal reproduction; (4) mutation of all neutral markers with a mutation rate of 10^{-3}; (5) migration of all individuals from A towards R and S. A sampling takes place every year at the end of generation 9 on S and R and at generation 11 before the migration event (5) on A.



(a) Complete summary statistics

(b) Wrap-up summary statistics

Figure 2: Principal component analyses (95% envelope) of simulations for each model: 'with' and 'without' host alternation, depending on the summary statistics considered. PCA analyses were based on the 'full' simulation design.

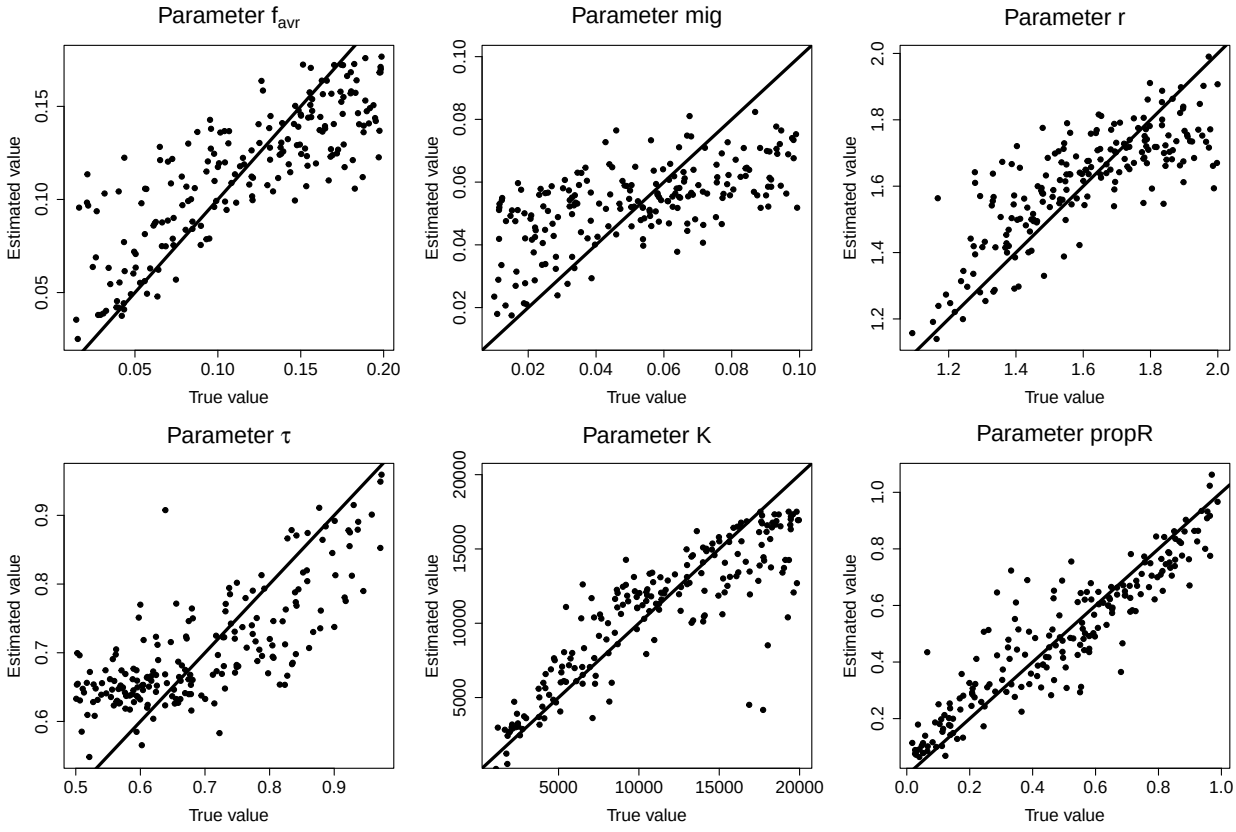
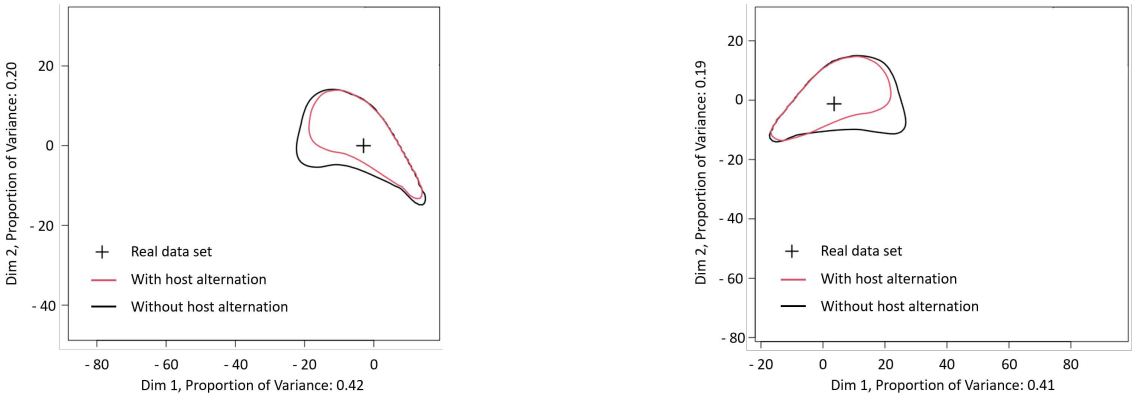


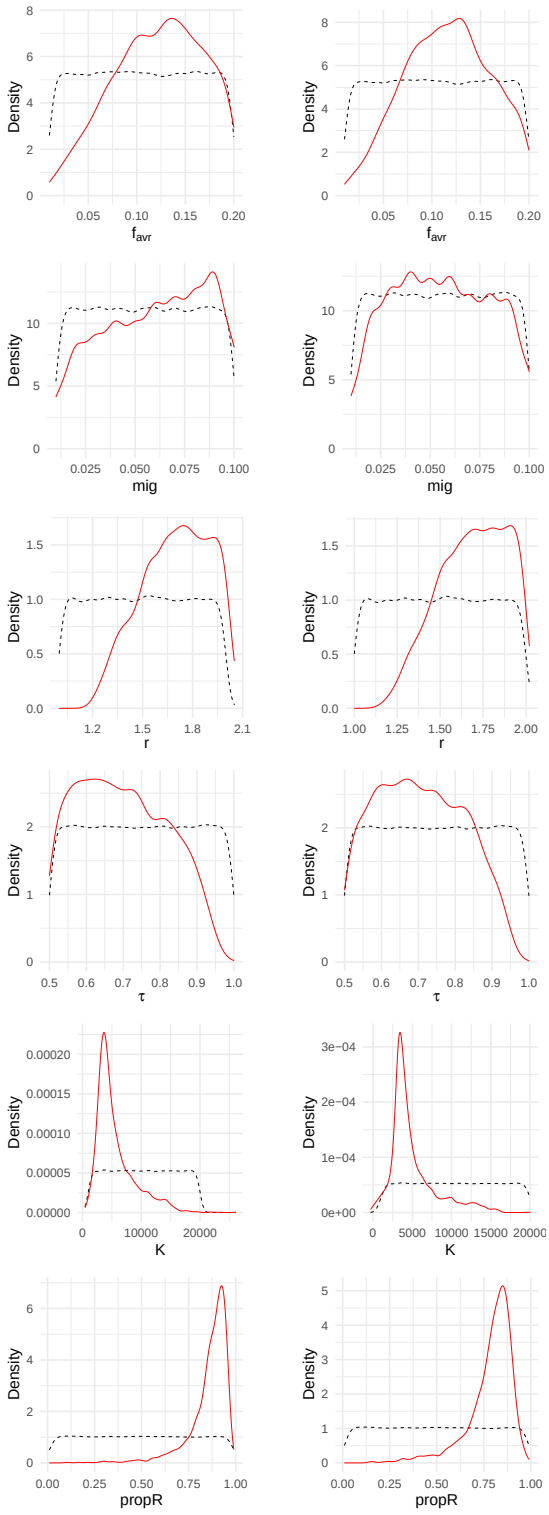
Figure 3: Practical identifiability of parameter estimation for Wrap-up summary statistics. Each point represents the parameter estimation ('Estimated' value) depending on the real parameter ('True' value). Each graph regroups the results of 200 replicates. Straight lines correspond to the first bisector.



(a) Amance

(b) Grand Est region

Figure 4: Principal component analyses (95% envelope) of simulations for each model: 'with' and 'without' host alternation, depending on the data set considered. Simulations were performed with Complete summary statistics. The black crosses correspond to the coordinates of the actual data sets.



(a) Amance (b) Grand Est region

Figure 5: Posterior distributions of parameters, with Complete summary statistics, for the data sets in Amance location (on S and A) and the Grand Est region (on S, R and A). Dashed lines correspond to the prior distributions and red lines correspond to the posterior distributions given by the neuralnet method.