Likelihood-Free Inference in High-Dimensional Models

Athanasios Kousathanas,*,† Christoph Leuenberger,‡ Jonas Helfer,§ Mathieu Quinodoz, Matthieu Foll,†† and Daniel Wegmann*,†,1**

*Department of Biology and Biochemistry and ‡Department of Mathematics, University of Fribourg, 1700 Fribourg, Switzerland, †Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland, §Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge Massachusetts 02139, **Department of Computational Biology, University of Lausanne, 1200 Lausanne, Switzerland, and ††International Agency for Research on Cancer, 69372 Lyon, France

ABSTRACT Methods that bypass analytical evaluations of the likelihood function have become an indispensable tool for statistical inference in many fields of science. These so-called likelihood-free methods rely on accepting and rejecting simulations based on summary statistics, which limits them to low-dimensional models for which the value of the likelihood is large enough to result in manageable acceptance rates. To get around these issues, we introduce a novel, likelihood-free Markov chain Monte Carlo (MCMC) method combining two key innovations: updating only one parameter per iteration and accepting or rejecting this update based on subsets of statistics approximately sufficient for this parameter. This increases acceptance rates dramatically, rendering this approach suitable even for models of very high dimensionality. We further derive that for linear models, a one-dimensional combination of statistics per parameter is sufficient and can be found empirically with simulations. Finally, we demonstrate that our method readily scales to models of very high dimensionality, using toy models as well as by jointly inferring the effective population size, the distribution of fitness effects (DFE) of segregating mutations, and selection coefficients for each locus from data of a recent experiment on the evolution of drug resistance in influenza.

KEYWORDS approximate Bayesian computation; distribution of fitness effects; hierarchical models; high dimensions; Markov chain Monte Carlo

THE past decade has seen a rise in the application of Bayesian inference algorithms that bypass likelihood calculations with simulations. Indeed, these generally termed likelihood-free or approximate Bayesian computation (ABC) (Beaumont *et al.* 2002) methods have been applied in a wide range of scientific disciplines, including cosmology (Schafer and Freeman 2012), ecology (Jabot and Chave 2009, protein-network evolution (Ratmann *et al.* 2007), phylogenetics (Fan and Kubatko 2011), and population genetics (Cornuet *et al.* 2008). Arguably ABC has had its greatest success in population genetics because inferences in this field are frequently conducted under complex models for which likelihood calculations are intractable, thus necessitating inference through simulations.

Let us consider a model $\mathcal{M}$ that depends on n parameters $\boldsymbol{\theta}$, creates data D, and has the posterior distribution

$$\pi(\boldsymbol{\theta}|D) = \frac{\mathbb{P}(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \mathbb{P}(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where $\pi(\boldsymbol{\theta})$ is the prior and $\mathbb{P}(D|\boldsymbol{\theta})$ is the likelihood function. ABC methods bypass the evaluation of $\mathbb{P}(D|\boldsymbol{\theta})$ by performing simulations with parameter values sampled from $\pi(\boldsymbol{\theta})$ that generate D, which in turn is summarized by a set of m-dimensional statistics $\boldsymbol{s}$. The posterior distribution is then evaluated by accepting such simulations that reproduce the statistics calculated from the observed data ($\boldsymbol{s}_{\text{obs}}$)

$$\pi(\boldsymbol{\theta}|\boldsymbol{s}) = \frac{\mathbb{P}(\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \mathbb{P}(\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

However, for models with $m \gg 1$ the condition $\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}$ might be too restrictive and require a prohibitively large simulation effort. Therefore, an approximation step can be employed by relaxing the condition $\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}$ to $\|\boldsymbol{s}-\boldsymbol{s}_{\text{obs}}\| \leq \delta$, where

Supplemental material is available online

[1] Corresponding author: Department of Biology, University of Fribourg, Chemin du Musée 10, 1200 Fribourg, Switzerland. E-mail: daniel.wegmann@unifr.ch

$\|x-y\|$ is a distance metric of choice between x and y and δ is a chosen distance (tolerance) below which simulations are accepted. The posterior $\pi(\boldsymbol{\theta}|\boldsymbol{s})$ is thus approximated by

$$\pi(\boldsymbol{\theta}|\boldsymbol{s}) = \frac{\mathbb{P}(\|\boldsymbol{s}-\boldsymbol{s}_{\text{obs}}\| \leq \delta|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \mathbb{P}(\|\boldsymbol{s}-\boldsymbol{s}_{\text{obs}}\| \leq \delta|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

An important advance in ABC inference was the development of methods coupling ABC with Markov chain Monte Carlo (MCMC) (Marjoram *et al.* 2003). These methods allow efficient sampling of the parameter space in regions of high likelihood, thus requiring fewer simulations to obtain posterior estimates (Wegmann *et al.* 2009). The original ABC-MCMC algorithm proposed by Marjoram *et al.* (2003) is as follows:

1. If now at $\boldsymbol{\theta}$ propose to move to $\boldsymbol{\theta}'$ according to the transition kernel $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
2. Simulate D using model $\mathcal{M}$ with $\boldsymbol{\theta}'$ and calculate summary statistics $\boldsymbol{s}$ for D.
3. If $\|\boldsymbol{s}-\boldsymbol{s}_{\text{obs}}\| \leq \delta$, go to step 4; otherwise go to step 1.
4. Calculate the Metropolis–Hastings ratio

$$h = h\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \min\left(1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}')}\right).$$

5. Accept $\boldsymbol{\theta}'$ with probability h; otherwise stay at $\boldsymbol{\theta}$. Go to step 1.

The sampling success of ABC algorithms is given by the likelihood values, which are often very low even for relatively large tolerance values δ. In such situations, the condition $\|\boldsymbol{s}-\boldsymbol{s}_{\text{obs}}\| \leq \delta$ will impose a quite rough approximation to the posterior. As a result, the utility of the ABC approaches described above is limited to models of relatively low dimensionality, typically up to 10 parameters (Blum 2010; Fearnhead and Prangle 2012). The same limitation applies to the more recently developed sequential Monte Carlo sampling methods (Sisson *et al.* 2007; Beaumont *et al.* 2009). Despite these limitations ABC has been useful in addressing population genetics problems of low to moderate dimensionality such as the inference of demographic histories (*e.g.*, Wegmann and Excoffier 2010; Brown *et al.* 2011; Adrion *et al.* 2014) or selection coefficients of a single locus (*e.g.*, Jensen *et al.* 2008). However, as more genomic data become available, there is increasing interest in applying ABC to models of higher dimensionality, such as to estimate genome-wide and locus-specific effects jointly.

To our knowledge, to date, three approaches have been suggested to tackle high dimensionality with ABC. The first approach proposes an expectation propagation approximation to factorize the data space (Barthelmé and Chopin 2014), which is an efficient solution for situations with high-dimensional data, but does not directly address the issue of high-dimensional parameter spaces. The second approach consists of first inferring marginal posterior distributions on low-dimensional subsets of the parameter space [either one (Nott *et al.* 2012) or two dimensions (Li *et al.* 2015)] and then reconstructing the joint posterior distribution from those. This approach benefits from the lower dimensionality of the statistics space when considering subsets of the parameters individually and hence renders the acceptance criterion meaningful. The third approach achieves the same benefit by formulating the problem using hierarchical models, proposing to estimate the hyperparameters first, and then fixing them when inferring parameters of lower hierarchies individually (Bazin *et al.* 2010).

Among these, the approach by Bazin *et al.* (2010) is the most relevant for population genetics problems, since those are frequently specified in a hierarchical fashion by modeling genome-wide effects as hyperparameters and locus-specific effects at lower hierarchies. In this way, Bazin *et al.* (2010) estimated locus-specific selection coefficients and deme-specific migration rates of an island model from microsatellite data. Furthermore, this approach has inspired the development of similar methods for estimating more complex migration patterns (Aeschbacher *et al.* 2013) and locus-specific selection from time-series data (Foll *et al.* 2015). However, this approach and its derivatives will not recover the true joint distribution if parameters are correlated, which is a common feature of such complex models.

Here, we introduce a new ABC algorithm that exploits the reduction of dimensionality of the summary statistics when focusing on subsets of parameters, but couples the parameter updates in an MCMC framework. As we prove below, this coupling ensures that our algorithm converges to the true joint posterior distribution even for models of very high dimensions. We then demonstrate its usefulness by inferring the effective population size jointly with locus-specific selection coefficients and the hierarchical parameters of the distribution of fitness effects (DFE) from allele frequency time-series data.

Theory

Let us define the random variable $\boldsymbol{T}_i = \boldsymbol{T}_i(\boldsymbol{s})$ as an m_i-dimensional function of $\boldsymbol{s}$. We call $\boldsymbol{T}_i$ *sufficient* for the parameter θ_i if the conditional distribution of $\boldsymbol{s}$ given $\boldsymbol{T}_i$ does not depend on θ_i. More precisely, let $\boldsymbol{t}_{i,\text{obs}} = \boldsymbol{T}_i(\boldsymbol{s}_{\text{obs}})$. Then

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}|\boldsymbol{T}_i=\boldsymbol{t}_{i,\text{obs}},\boldsymbol{\theta}\right) &= \frac{\mathbb{P}\left(\boldsymbol{s}=\boldsymbol{s}_{\text{obs}},\boldsymbol{T}_i=\boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}\right)}{\mathbb{P}\left(\boldsymbol{T}_i=\boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}\right)} \\ &= \frac{\mathbb{P}\left(\boldsymbol{s}=\boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}\right)}{\mathbb{P}\left(\boldsymbol{T}_i=\boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}\right)} \\ &=: g_i(\boldsymbol{s}_{\text{obs}},\boldsymbol{\theta}_{-i}), \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ is $\boldsymbol{\theta}$ with the ith component omitted.

It is not hard to find examples for parameter-wise sufficient statistics. Most common distributions are members of the exponential family, and for these, the density of $\boldsymbol{s}$ has the form

$$f(\boldsymbol{s}|\boldsymbol{\theta}) = h(\boldsymbol{s})\exp\left[\sum_{k=1}^{K}\eta_k(\boldsymbol{\theta})T_k(\boldsymbol{s}) - A(\boldsymbol{\theta})\right].$$

For a given parameter θ_i, the vector $\boldsymbol{T}_i(\boldsymbol{s})$ consisting of only those $T_k(\boldsymbol{s})$ for which the respective natural parameter function $\eta_k(\boldsymbol{\theta})$ depends on θ_i is a sufficient statistic for θ_i in the sense of our definition. Some concrete examples of this type are studied below.

If sufficient statistics $\boldsymbol{T}_i$ can be found for each parameter θ_i and their dimension m_i is substantially smaller than the dimension m of $\boldsymbol{s}$, then the ABC-MCMC algorithm can be greatly improved with the following algorithm that we denote ABC with **pa**rameter-**s**pecific **s**tatistics (ABC-PaSS) henceforth.

The algorithm starts at time $t = 1$ and at some initial parameter value $\boldsymbol{\theta}^{(1)}$.

1. Choose an index $i = 1, \dots, n$ according to a probability distribution $(p_1, \dots, p_n)$ with $\sum p_i = 1$ and all $p_i > 0$.
2. At $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ propose $\boldsymbol{\theta}'$ according to the transition kernel $q_i(\boldsymbol{\theta}'|\boldsymbol{\theta})$ where $\boldsymbol{\theta}'$ differs from $\boldsymbol{\theta}$ only in the ith component:
$$\boldsymbol{\theta}' = (\theta_1, \dots, \theta_{i-1}, \theta_i', \theta_{i+1}, \dots, \theta_n).$$
3. Simulate D using model $\mathcal{M}$ with $\boldsymbol{\theta}'$ and calculate summary statistics $\boldsymbol{s}$ for D. Calculate $\boldsymbol{t}_i = \boldsymbol{T}_i(\boldsymbol{s})$ and $\boldsymbol{t}_{i,\text{obs}} = \boldsymbol{T}_i(\boldsymbol{s}_{\text{obs}})$.
4. Let δ_i be the tolerance for parameter $\boldsymbol{\theta}_i$. If $\left|\left|\boldsymbol{t}_i - \boldsymbol{t}_{i,\text{obs}}\right|\right|_i \leq \delta_i$, go to step 5; otherwise go to step 1.
5. Calculate the Metropolis–Hastings ratio
$$h = h\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \min\left(1, \frac{\pi(\boldsymbol{\theta}')q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right)}{\pi(\boldsymbol{\theta})q_i\left(\boldsymbol{\theta}'|\boldsymbol{\theta}\right)}\right).$$
6. Accept $\boldsymbol{\theta}'$ with probability h; otherwise stay at $\boldsymbol{\theta}$.
7. Increase t by one, save a new parameter value $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$, and continue at step 1.

Convergence of the MCMC chain is guaranteed by the following:

Theorem 1. *For $i = 1..n$, if $\delta_i = 0$ and $\boldsymbol{T}_i$ is sufficient for parameter θ_i, then the stationary distribution of the Markov chain is $\pi(\boldsymbol{\theta}|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})$.*

The *Proof for Theorem 1* is provided in the *Appendix*.

It is important to note that the same algorithm can also be applied to groups of parameters, which may be particularly relevant in the case of very high correlations between parameters that may render their individual MCMC updates inefficient. Also, the efficiency of ABC-PaSS can be improved with all previously proposed extensions for ABC-MCMC. To increase acceptance rates and render ABC-PaSS applicable to models with continuous sampling distributions, for instance, the assumption $\delta_i = 0$ must be relaxed to $\delta_i > 0$ in practice. This is commonly done in ABC applications and will lead to an approximation of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})$. Because of the continuity of the summary statistics s and the sufficient statistics $\boldsymbol{T}_i$, we theoretically recover the true posterior distribution in the limit $\delta_i \rightarrow 0$. We can also perform an initial calibration ABC step to find an optimal starting position $\boldsymbol{\theta}^{(1)}$ and tolerance δ_i and to adjust the proposal kernel for each parameter (Wegmann *et al.* 2009).

Materials and Methods

Implementation

We implemented the proposed ABC-PaSS framework into a new version of the software package ABCtoolbox (Wegmann *et al.* 2010), which will be made available at the authors' website and will be described elsewhere.

Toy model 1: Normal distribution

We performed simulations to assess the performance of ABC-MCMC and ABC-PaSS in estimating $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ for a univariate normal distribution. We used the sample mean $\overline{x}$ and sample variance S^2 of samples of size n as statistics. Recall that for noninformative priors the posterior distribution for μ is $\mathcal{N}(\overline{x}, S^2/n)$ and the posterior distribution for σ^2 is χ^2 distributed with $n-1$ d.f. As μ and σ^2 are independent, we get the posterior density

$$\pi\left(\mu, \sigma^2\right) = \phi_{\overline{x}, S^2/n}(\mu) \cdot \frac{n-1}{S^2} f_{\chi^2;n-1}\left(\frac{n-1}{S^2}\sigma^2\right).$$

In our simulations the sample size was $n = 10$ and the true parameters were given by $\mu = 0$ and $\sigma^2 = 5$. We performed 50 MCMC chains per simulation and chose effectively noninformative priors for $\mu \sim U[-10, 10]$ and $\sigma^2 \sim U[0.1, 15]$. Our simulations were performed for a wide range of tolerances (from 0.01 to 41) and proposal ranges (from 0.05 to 1.5). We did this exhaustive search to identify the combination of these tuning parameters that allows ABC-MCMC and ABC-PaSS to perform best in estimating μ and σ^2. We then recorded the minimum total variation distance (L_1) between the true and estimated posteriors over these sets of tolerances and ranges and compared it between ABC-MCMC and ABC-PaSS.

Toy model 2: General linear model

As a second toy model to compare the performance of ABC-MCMC and ABC-PaSS, we considered general linear models (GLMs) with m statistics $\boldsymbol{s}$ being a linear function of $n = m$ parameters $\boldsymbol{\theta}$,

$$\boldsymbol{s} = \boldsymbol{C}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

where $\boldsymbol{C}$ is a square design matrix and the vector of errors $\boldsymbol{\epsilon}$ is multivariate normal. Under noninformative priors for the parameters $\boldsymbol{\theta}$, their posterior distribution is multivariate normal

$$\boldsymbol{\theta}|\boldsymbol{s} \sim \mathcal{N}\left(\left(\boldsymbol{C}'\boldsymbol{C}\right)^{-1}\boldsymbol{C}'\boldsymbol{s}, \left(\boldsymbol{C}'\boldsymbol{C}\right)^{-1}\right).$$

We set up the design matrices $\boldsymbol{C}$ in a cyclic manner to allow all statistics to have information on all parameters but their contributions to differ for each parameter; namely we set $\boldsymbol{C} = \mathbf{B} \cdot \det(\mathbf{B}'\mathbf{B})^{-1/2n}$, where

$$\mathbf{B} = \begin{pmatrix} 1/n & 2/n & 3/n & \dots & n/n \\ n/n & 1/n & 2/n & \dots & n-1/n \\ \vdots & \vdots & \vdots & \ddots & 2/n \\ 2/n & 3/n & 4/n & \dots & 1/n \end{pmatrix}.$$

The normalization factor in the definition of $\boldsymbol{C}$ was chosen such that the determinant of the posterior variance is constant and thus the widths of the marginal posteriors are comparable independently of the dimensionality n. We used all statistics for ABC-MCMC and calculated a single linear combination of statistics per parameter for ABC-PaSS according to *Theorem 2*, using ordinary least squares. For the estimation, we assumed that $\boldsymbol{\theta} = \mathbf{0}$ and the priors are uniform $U[-100, 100]$ for all parameters, which are effectively noninformative. We started the MCMC chains at a normal deviate $N(\boldsymbol{\theta}, 0.01\boldsymbol{I})$, *i.e.*, around the true values of $\boldsymbol{\theta}$. To ensure fair comparisons between methods, we performed simulations of 50 chains for a variety of tolerances (from 0.01 to 256) and proposal ranges (from 0.1 to 8) to choose the combination of these tuning parameters at which each method performed best. We ran all our MCMC chains for 10^5 iterations per model parameter to account for model complexity.

Estimating selection and demography

Model: Consider a vector $\boldsymbol{\xi}$ of observed allele trajectories (sample allele frequencies) over $l = 1, \dots, L$ loci, as is commonly obtained in studies of experimental evolution. We assume these trajectories to be the result of both random drift and selection, parameterized by the effective population size N_e and locus-specific selection coefficients s_l, respectively, under the classic Wright–Fisher model with allelic fitnesses 1 and $1 + s_l$. We further assume the locus-specific selection coefficients s_l follow a DFE parameterized as a generalized Pareto distribution (GPD) with mean $\mu = 0$, shape χ, and scale σ. Our goal is thus to estimate the joint posterior distribution

$$\begin{aligned}&\pi(N_\mathrm{e}, s_1, \dots, s_L, \chi, \sigma|\boldsymbol{\xi})\\ &\propto \prod_{l=1}^{L}\left[\mathbb{P}(\xi_l|N_\mathrm{e}, s_l)\pi(s_l|\chi, \sigma)\right]\pi(N_\mathrm{e})\pi(\chi)\pi(\sigma).\end{aligned}$$

To apply our ABC-PaSS framework to this problem, we approximate the likelihood term $\mathbb{P}(\xi_l|N_\mathrm{e}, s_l)$ numerically with simulations, while updating the hyperparameters χ and σ analytically.

Summary statistics: To summarize the data $\boldsymbol{\xi}$, we used statistics originally proposed by Foll *et al.* (2015). Specifically, we first calculated for each locus individually a measure of the difference in allele frequency between consecutive time points as

$$Fs' = \frac{1}{t}\,\frac{Fs[1 - 1/(2\tilde{n})] - 2/\tilde{n}}{(1 + Fs/4)[1 - 1/(n_y)]},$$

where

$$Fs = \frac{(x-y)^2}{z(1-z)},$$

x and y are the minor allele frequencies separated by t generations, $z = (x + y)/2$, and $\tilde{n}$ is the harmonic mean of the sample sizes n_x and n_y. We then summed the Fs' values of all pairs of consecutive time points with increasing and decreasing allele frequencies into $Fs'i$ and $Fs'd$, respectively (Foll *et al.* 2015). Finally, we followed Aeschbacher *et al.* (2012) and calculated boosted variants of the two statistics to take more complex relationships between parameters and statistics into account. The full set of statistics used per locus was $\boldsymbol{F}_l = \{Fs'i_l, Fs'd_l, Fs'i_l^2, Fs'd_l^2, Fs'i_l \times Fs'd_l\}$

We next calculated parameter-specific linear combinations for N_e and locus-specific s_l following the procedure developed above. To do so, we simulated allele trajectories of a single locus for different values of N_e and s sampled from their prior. We then calculated $\boldsymbol{F}_l$ for each simulation and performed a Box–Cox transformation to linearize the relationships between statistics and parameters (Box and Cox 1964; Wegmann *et al.* 2009). We then fitted a linear model as outlined in Equation A3 to estimate the coefficients of an approximately sufficient linear combination of $\boldsymbol{F}$ for each parameter N_e and s. This resulted in $\tau_s(\boldsymbol{F}_l) = \boldsymbol{\beta}_s\boldsymbol{F}_l$ and $\tau_{N_\mathrm{e}}(\boldsymbol{F}_l) = \boldsymbol{\beta}_{N_\mathrm{e}}\boldsymbol{F}_l$. To combine information across loci when updating N_e, we then calculated

$$\tau_{N_\mathrm{e}}(\boldsymbol{F}) = \sum_{l=1}^{L}\boldsymbol{\beta}_{N_\mathrm{e}}\boldsymbol{F}_l,$$

where $\boldsymbol{F} = \{\boldsymbol{F}_1, \dots, \boldsymbol{F}_L\}$ In summary, we used the ABC approximation

$$\begin{aligned}\mathbb{P}(\xi_j|N_\mathrm{e}, s_j) \approx \mathbb{P}\big(\|\tau_s(\boldsymbol{F}_l) - \tau_s(\boldsymbol{F}_{l_\mathrm{obs}})\| < \delta_{s_l},\\ \|\tau_{N_\mathrm{e}}(\boldsymbol{F}) - \tau_{N_\mathrm{e}}(\boldsymbol{F}_\mathrm{obs})\| < \delta_{N_\mathrm{e}}|N_\mathrm{e}, s_j\big).\end{aligned}$$

Simulations and application

We applied our framework to allele frequency data for the whole influenza H1N1 genome obtained in a recently published evolutionary experiment (Foll *et al.* 2014). In this experiment, influenza A/Brisbane/59/2007 (H1N1) was serially amplified on Madin–Darby canine kidney (MDCK) cells for 12 passages of 72 hr each, corresponding to ~13 generations (doublings). After the three initial passages, samples were passed either in the absence of drug or in the presence of increasing concentrations of the antiviral drug oseltamivir. At the end of each passage, samples were collected for whole-genome high-throughput population sequencing. We obtained the raw data from http://bib.umassmed.edu/influenza/ and, following the original study (Foll *et al.* 2014), we downsampled it to 1000 haplotypes per time point and filtered it to contain only loci for which sufficient data were available to calculate the Fs' statistics. Specifically, we included all loci

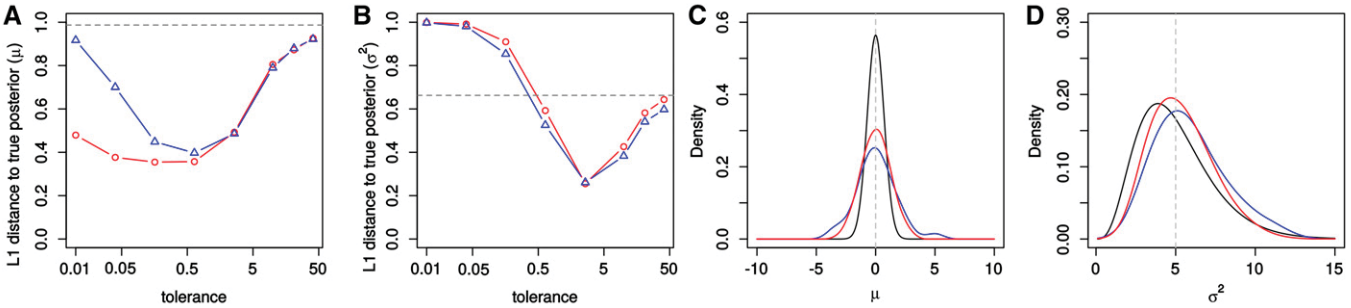

Figure 1 Performance to infer parameters of a normal distribution. Shown is the average over 50 chains of the L_1 distance between the true and estimated posterior distributions for μ (A) and σ^2 (B) for different tolerances for ABC-MCMC (blue) and ABC-PaSS (red). The dashed horizontal line is the L_1 distance between the prior and the true posterior distribution. (C and D) The estimated posterior distribution for μ (C) and σ^2 (D) using the tolerance that led to the minimum L_1 distance from the true posterior (black). The dashed vertical line indicates the true values of the parameters.

with an allele frequency $\geq$ 2% at $\geq$ 2 time points. There were 86 and 42 such loci for the control and drug-treated experiments, respectively. Further, we restricted our analysis of the data of the drug-treated experiment to the last nine time points during which drug was administered.

We performed all our Wright–Fisher simulations with in-house C++ code implemented as a module of ABCtoolbox. We simulated 13 generations between time points and a sample of size 1000 per time point. We set the prior for N_e uniform on the $\log_{10}$ scale such that $\log_{10}(N_e) \sim U[1.5, 4.5]$ and for the parameters of the GPD $\chi \sim U[-0.2, 1]$ and for $\log_{10}(\sigma) \sim U[-2.5, -0.5]$. For the simulations where no DFE was assumed, we set the prior of $s \sim U[0, 1]$.

As above, we ran all our ABC-PaSS chains for 10^5 iterations per model parameter to account for model complexity. To ensure fast convergence, the ABC-PaSS implementation benefited from an initial calibration step we originally developed for ABC-MCMC and implemented in ABCtoolbox (Wegmann *et al.* 2009). Specifically, we first generated 10,000 simulations with values drawn randomly from the prior. For each parameter, we then selected the 1% subset of these simulations with the smallest distances to the observed data based on the linear combination specific for that parameter. These accepted simulations were used to calibrate three important metrics prior to the MCMC run: First, we set the parameter-specific tolerances δ_i to the largest distance among the accepted simulations. Second, we set the width of the parameter-specific proposal kernel to half of the standard deviation of the accepted parameter values. Third, we chose the starting value of the chain for each parameter as the accepted simulation with smallest distance. Each chain was then run for 1000 iterations, and new starting values were chosen randomly among the accepted calibration simulations for those parameters for which no update was accepted. This was repeated until all parameters were updated at least once.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

Toy model 1: Normal distribution

We first compared the performance of ABC-PaSS and ABC-MCMC under a simple model: the normal distribution with parameters mean (μ) and variance (σ^2). Given a sample of size n, the sample mean ($\bar{x}$) is a sufficient statistic for μ, while both $\bar{x}$ and the sample variance (S^2) are sufficient for σ^2 (Casella and Berger 2002). For ABC-MCMC, we used both $\bar{x}$ and S^2 as statistics. For ABC-PaSS, we used only $\bar{x}$ when updating μ and both $\bar{x}$ and S^2 when updating σ^2.

We then compared the accuracy between the two algorithms by calculating the total variation distance between the inferred and the true posteriors (L_1 distance from kernel smoothed posterior based on 10,000 samples). We computed L_1 under a wide range of tolerances to find the tolerance for which each algorithm had the best performance (*i.e.*, minimum L_1). As shown in Figure 1, A and C, ABC-PaSS produced a more accurate estimation for μ than ABC-MCMC. The two algorithms had similar performance when estimating σ^2 (Figure 1, B and D).

The normal distribution toy model, although simple, is quite illustrative of the nature of the improvement in performance by using ABC-PaSS over ABC-MCMC. Indeed, our results demonstrate that the slight reduction of the summary statistics space by ignoring a single uninformative statistic when updating μ already results in a noticeable improvement in estimation accuracy. This improvement would not be possible to attain with classic dimension reduction techniques, such as partial least squares (PLS), since the information contained in $\bar{x}$ and S^2 is irreducible under ABC-MCMC.

Toy model 2: GLM

We expect our approach to be particularly powerful for models of the exponential family, for which a small number of summary statistics per parameter are sufficient, regardless of sample size. To illustrate this, we next compared the performance of ABC-MCMC and ABC-PaSS under GLM models of increasing dimensionality n. For all models, we constructed the design matrix $\boldsymbol{C}$ such that all statistics are informative for all parameters, while retaining the total information on the

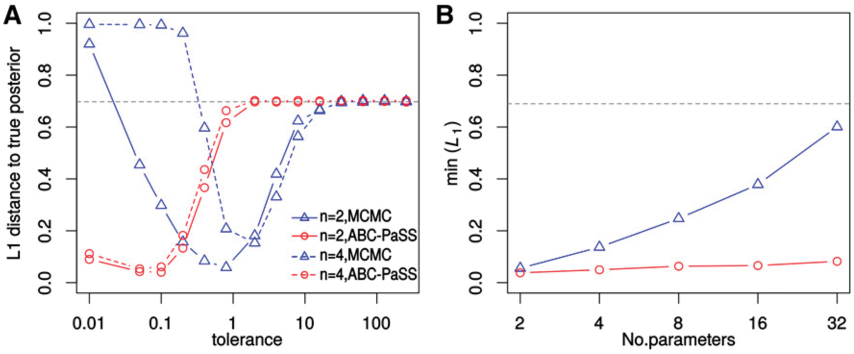

Figure 2 Performance to infer parameters of GLM models. (A) The average L_1 distance between the true and estimated posterior distributions for different tolerances for ABC-MCMC (blue) and ABC-PaSS (red). Solid and dashed lines are for a GLM with two and four parameters, respectively. (B) The minimum L_1 distance from the true posterior over different tolerances for increasing numbers of parameters. (A and B) The dashed line is the L_1 distance between the prior and the posterior distribution.

individual parameters regardless of dimensionality (see *Materials and Methods*). For a GLM, a single linear function is a sufficient statistic for each associated parameter, and this function can easily be learned from a set of simulations, using standard regression approaches (see *Theorem 2* in the *Appendix*). Therefore, for ABC-MCMC, we used all statistics $\boldsymbol{s}$, while for ABC-PaSS, we employed *Theorem 2* and used a single linear combination of statistics τ_i per parameter θ_i. As above, we assessed performance of ABC-MCMC and ABC-PaSS by calculating the total variation distance (L_1) between the inferred and the true posterior distribution. We calculated L_1 for several tolerances to find the tolerance where L_1 was minimal for each algorithm (see Figure 2A for examples with $n = 2$ and $n = 4$). Since in ABC-MCMC distances are calculated in the multidimensional statistics space, the optimal tolerance increased with higher dimensionality. This is not the case for ABC-PaSS, because distances are always calculated in one dimension only (Figure 2A).

We found that ABC-MCMC performance was good for low n, but worsened rapidly with increasing number of parameters, as expected from the corresponding increase in the dimensionality of statistics space (Figure 2B). For a GLM with 32 parameters, approximate posteriors obtained with ABC-MCMC differed only little from the prior (Figure 2B). In contrast, performance of ABC-PaSS was unaffected by dimensionality and was better than that of ABC-MCMC even in low dimensions (Figure 2B). These results support that by considering low-dimensional parameter-specific summary statistics under our framework, ABC inference remains feasible even under models of very high dimensionality, for which current ABC algorithms are not capable of producing meaningful estimates.

Application: Inference of natural selection and demography

One of the major research problems in modern population genetics is the inference of natural selection and demographic history, ideally jointly (Crisci *et al.* 2012; Bank *et al.* 2014). One way to gain insight into these processes is by investigating how they affect allele frequency trajectories through time in populations, for instance under experimental evolution. Several methods have thus been developed to analyze allele trajectory data to infer both locus-specific selection coefficients (s) and the effective population size (N_e). The modeling framework of these methods assumes Wright–Fisher (WF) population dynamics in a hidden Markov setting to evaluate the likelihood of the parameters N_e and s given the observed allele trajectories (Bollback *et al.* 2008; Malaspinas *et al.* 2012). In this setting, likelihood calculations are feasible, but very time-consuming, especially when considering many loci at the genome-wide scale (Foll *et al.* 2015).

To speed up calculations, Foll *et al.* (2015) developed an ABC method (WF-ABC), adopting the hierarchical ABC framework of Bazin *et al.* (2010). Specifically, WF-ABC first estimates N_e based on statistics that are functions of all loci and then infers s for each locus individually under the inferred value of N_e. While WF-ABC easily scales to genome-wide data, it suffers from the unrealistic assumption of complete neutrality when inferring N_e, which potentially leads to biases in the inference.

Here we show that by employing ABC-PaSS, N_e and locus-specific selection coefficients can be inferred jointly, which is not possible with ABC-MCMC due to high dimensionality of the summary statistics that is a direct function of the number of loci considered.

Finding sufficient statistics: All ABC algorithms, including ABC-PaSS introduced here, require that statistics are sufficient for estimating the parameters of a given model. As mentioned above, parameter-wise sufficient statistics as required by ABC-PaSS are trivial to find for distributions of the exponential family. Since many population genetics models do not follow such distributions, sufficient statistics are known for the most simple models only. The number of haplotypes segregating in a sample, for example, is a sufficient statistic for estimating the population-scaled mutation rate under Wright–Fisher equilibrium assumptions (Durrett 2008).

For more realistic models involving multiple populations or population size changes, only approximately-sufficient statistics can be found. Choosing such statistics is not trivial, however, as too few statistics are insufficient to summarize the data while too many statistics can create an excessively large statistics space that worsens the approximation of the posterior (Beaumont *et al.* 2002; Wegmann *et al.* 2009; Csilléry *et al.* 2010). Often, such statistics are thus found

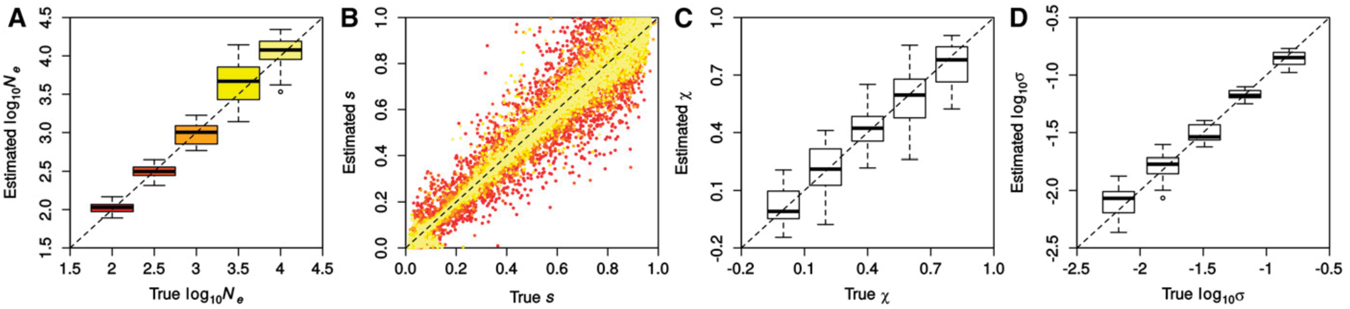

Figure 3 Accuracy in inferring demographic and selection parameters. Results were obtained with ABC-PaSS using a single combination of statistics for N_e and each s (LC 1/1). Shown are the true *vs.* estimated posterior medians for parameters N_e (A), s per locus (B), and χ and σ of the generalized Pareto distribution (C and D, respectively). Boxplots summarize results from 25 replicate simulations, each with 100 loci. Uniform priors over the whole ranges shown were used. (A and B) N_e assumed in the simulations is represented as a color gradient of red (low N_e) to yellow (high N_e). (C and D) Parameters μ and N_e were fixed to 0 and 10^3, respectively; $\log_{10}\sigma$ was fixed to -1 (C); and χ was fixed to 0.5 (D).

empirically by applying dimensionality reduction techniques to a larger set of statistics initially calculated (Blum *et al.* 2013).

Fearnhead and Prangle (2012) suggested a method where an initial set of simulations is used to fit a linear model, using ordinary least squares that expresses each parameter θ_i as a function of the summary statistics $\boldsymbol{s}$. These functions are then used as statistics in subsequent ABC analysis. Thus Fearnhead and Prangle's approach reduces the dimensionality of statistics space to a single combination of statistics per parameter. However, the Pitman–Koopman–Darmois theorem states that for models that do not belong to the exponential family, the dimensionality of sufficient statistics must grow with increasing sample size, suggesting that multiple summary statistics are likely required in this case as any locus carries independent information for the parameter N_e. A method similar in spirit but not limited to a single summary statistic per parameter is a partial least-squares transformation (Wegmann *et al.* 2009), which has been used successfully in many ABC applications (*e.g.*, Veeramah *et al.* 2011; Chu *et al.* 2013; Dussex *et al.* 2014).

Here we chose to calculate the per locus statistics proposed by Foll *et al.* (2015) and to then apply and empirically compare both methods to reduce dimensionality for this particular model. Before dimension reduction, however, we applied a multivariate Box–Cox transformation (Box and Cox 1964) to increase linearity between statistics and parameters, as suggested by Wegmann *et al.* (2009). To decide on the required number of PLS components, we performed a leave-one-out analysis implemented in the R package "PLS" (Mevik and Wehrens 2007). In line with the Pitman–Koopman–Darmois theorem, a small number (two) of PLS components were sufficient for s, but many more components contained information about N_e, for which many independent observations are available (Supplemental Material, Figure S1). However, the first PLS component alone explained already two-thirds of the total variance than can be explained with up to 100 components, suggesting that additional components add, besides information, also substantial noise. We thus chose to evaluate the accuracy of our inference with three different sets of summary statistics: (1) a single linear combination of summary statistic for each s and N_e chosen using ordinary least squares, as suggested by Fearnhead and Prangle (2012) (LC 1/1); (2) two PLS components for s and five PLS components for N_e, as suggested by the leave-one-out analysis (PLS 5/2); and (3) an intermediate set of one PLS component for s and three PLS components for N_e (PLS 3/1).

Performance of ABC-PaSS in inferring selection and demography: To examine the performance of ABC-PaSS under the WF model, we inferred N_e and s on sets of 100 loci simulated with varying selection coefficients. We evaluated the estimation accuracy by comparing the estimated *vs.* the true values of the parameters over 25 replicate simulations, first using a single linear combination of summary statistics per parameter found using ordinary least squares (LC 1/1). As shown in Figure 3A, N_e was estimated well over the whole range of values tested. Estimates for s were on average unbiased and accuracy was, as expected, higher for larger N_e (Figure 3B). Note that since the prior on s was $U[0,1]$, these results imply that our approach estimates N_e with high accuracy even when the majority of the simulated loci are under strong selection (90% of loci had $N_e s > 10$). Hence, our method allows us to relax the assumption of neutrality on most of the loci, which was necessary in previous studies (Foll *et al.* 2015).

We next introduced hyperparameters for the distribution of selection coefficients (the so-called DFE). Such hyperparameters are computationally cheap to estimate under our framework, as their updates can be done analytically and do not require simulations. Following previous work (Beisel *et al.* 2007; Martin and Lenormand 2008), we assumed that the distribution of the locus-specific s is realistically described by a truncated GPD with location $\mu = 0$ and parameters shape σ and scale χ (Figure S2).

We first evaluated the accuracy of estimating χ and σ when fixing the value of the other parameter and found that both parameters are well estimated under these conditions (Figure 3, C and D, respectively). Since the truncated GPD of multiple combinations of χ and σ is very similar, these parameters are not always identifiable. This renders the

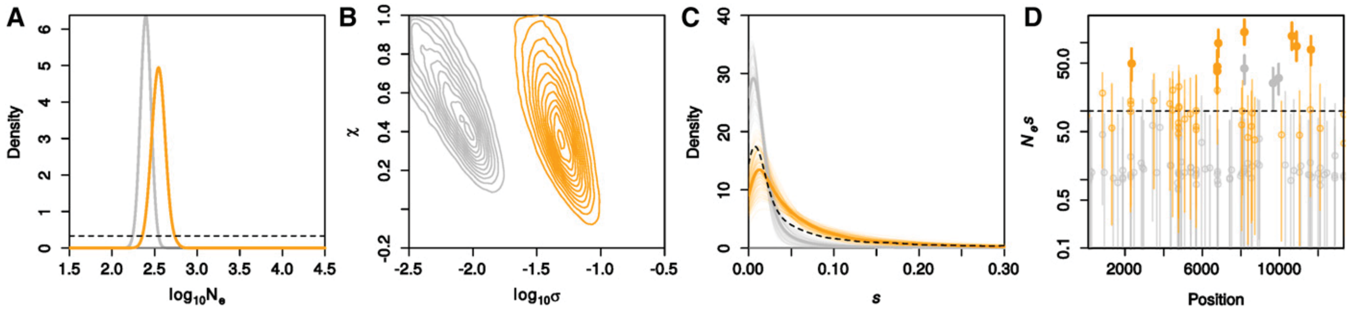

Figure 4 Inferred demography and selection for experimental evolution of influenza. We show results for the no-drug (control) and drug-treated influenza in gray and orange, respectively. Shown are the posterior distributions for $\log_{10} N_e$ (A) and $\log_{10}\sigma$ and χ (B). In C, we plotted the modal DFE with thick lines by integrating over the posterior of its parameters. The thin lines represent the DFEs obtained by drawing 100 samples from the posterior of σ and χ. Dashed lines in A and C correspond to the prior distributions. In D, the posterior estimates for $N_e s$ per locus *vs.* the position of the loci in the genome are shown. Open circles indicate nonsignificant loci whereas solid, thick circles indicate significant loci [$P(N_e s > 10) > 0.95$, dashed line].

accurate joint estimation of both parameters difficult (Figure S3, B and C). However, despite the reduced accuracy on the individual parameters, we found the overall shape of the GPD to be well recovered (Figure S3, D–F). Also, N_e was estimated with high accuracy for all combinations of χ and σ (Figure S3A).

We then checked whether the accuracy of these estimates can be improved by using summary statistics of higher dimensionality. Specifically, we repeated these analyses with a high-dimensional set (PLS 5/2) consisting of the first five and the first two PLS components for N_e and each s, respectively, as well as a set of intermediate dimensionality (PLS 3/1) consisting of the first three PLS components for N_e and only the first PLS component for each s. Overall, all sets of summary statistics compared here resulted in very similar performance as assessed both visually (compare Figure 3, Figure S4, and Figure S5 for LC 1/1, PLS 5/2, and PLS 3/1, respectively) and by calculating the both root mean square error (RMSE) and Pearson's correlation coefficient between true and inferred values (Table S1). Interestingly, the intermediate set (PLS 3/1) performed worst in all comparisons, while the differences between LC 1/1 and PLS 5/2 were very subtle, particularly when uniform priors were used on all s (simulation set 1; Table S1). However, in the presence of hyperparameters on s, results were more variable (simulation sets 2–4; Table S1) and we found the effective population size N_e to be consistently overestimated when using high-dimensional summaries such as PLS 5/2 (simulation sets 2–4; Table S1). These results suggest that while our analysis is generally rather robust to the choice of summary statistics, the benefit of extra information added by additional summary statistics is offset by the increased noise in higher dimensions. We expect that robustness of results to the choice of summary statistics will be model dependent and recommend that the performance of multiple-dimension reduction techniques should be evaluated in future applications of ABC-PaSS like we did here.

Analysis of infuenza data: We applied our approach to data from a previous study (Foll *et al.* 2014) where cultured canine kidney cells infected with the influenza virus were subjected to serial transfers for several generations. In one experiment, the cells were treated with the drug Oseltamivir, and in a control experiment they were not treated with the drug. To obtain allele frequency trajectories of all sites of the infuenza virus genome (13.5 kbp), samples were taken and sequenced every 13 generations with pooled population sequencing. The aim of our application was to identify which viral mutations rose in frequency during the experiment due to natural selection and which due to drift and to investigate the shape of the DFE for the control and drug-treated viral populations.

Following Foll *et al.* (2014), we filtered the raw data to contain loci for which sufficient data were available to calculate the summary statistics considered here (see *Materials and Methods*). There were 86 and 42 such loci for the control and drug-treated experiments, respectively (Figure S6).

We then employed ABC-PaSS to estimate N_e, s per locus and the parameters of the DFE, first using a single summary statistic per parameter (LC 1/1). We obtained a low estimate for N_e (posterior medians 350 for drug-treated and 250 for control influenza; Figure 4A), which is expected given the bottleneck that the cells were subjected to in each transfer. While we obtained similar estimates for the χ parameters for the drug-treated and for the control influenza (posterior medians 0.44 and 0.56, respectively), the σ parameter was estimated to be much higher for the drug-treated than for the control influenza (posterior medians 0.047 and 0.0071, respectively; Figure 4B). The resulting DFE was thus very different for the two conditions: The DFE for the drug-treated influenza had a much heavier tail than the control (Figure 4C). Posterior estimates for $N_e s$ per locus also indicated that the drug-treated influenza had more loci under strong positive selection than the control (19% *vs.* 3.5% of loci had $P(N_e s > 10) > 0.95$, respectively; Figure 4D and Figure S6). Almost identical results were also obtained when using higher-dimensional summary statistics based on PLS components (Figure S7). These results indicate that the drug treatment placed the influenza population away from a fitness optimum, thus increasing the number of positively selected mutations with large effect sizes. Presumably these mutations confer resistance to the drug, thus helping influenza to reach a new fitness optimum.

Our results for influenza were qualitatively similar to those obtained by Foll *et al.* (2014). We obtained slightly larger estimates for N_e (350 *vs.* 226 for drug-treated and 250 *vs.* 176 for control influenza). Our estimates for the parameters of the GPD were substantially different from those of Foll *et al.* (2014) but resulted in qualitatively similar overall shapes of the DFE for both drug-treated and control experiments. These results underline the applicability of our method to a high-dimensional problem. In contrast to Foll *et al.* (2014) who performed estimations in a three-step approach, combining a moment-based estimator for N_e, ABC for s, and a maximum-likelihood approach for the GPD, our Bayesian framework allowed us to perform joint estimation and to obtain posterior distributions for all parameters in a single step.

Discussion

Due to the difficulty to find analytically tractable likelihood solutions, statistical inference is often limited to models that made substantial approximations of reality. To address this problem, so-called likelihood-free approaches have been introduced that bypass the analytical evaluation of the likelihood function with computer simulations. While full-likelihood solutions generally have more power, likelihood-free methods have been used in many fields of science to overcome undesired model assumptions.

Here we developed and implemented a novel likelihood-free, MCMC inference framework that scales naturally to high dimensions. This framework takes advantage of the observation that the information about one model parameter is often contained in a subset of the data, by integrating two key innovations: First, only a single parameter is updated at a time, and the update is accepted based on a subset of summary statistics sufficient for this parameter. We proved that this MCMC variant converges to the true joint posterior distribution under the standard assumptions.

Since simulations are accepted based on lower dimensionality, our algorithm proposed here will have a higher acceptance rate than other ABC approaches for the same accuracy and hence require fewer simulations. This is particularly relevant for cases in which the simulation step is computationally challenging, such as for population genetic models that are spatially explicit (Ray *et al.* 2010) or require forward-in-time simulations (as opposed to coalescent simulations) (Hernandez 2008; Messer 2013).

We demonstrated the power of our framework through the application to multiple problems. First, our framework led to more accurate inference of the mean and standard deviation of a normal distribution than standard likelihood-free MCMC, suggesting that our framework is already competitive in models of low dimensionality. In high dimensions, the benefit was even more apparent. When applied to the problem of inferring parameters of a GLM, for instance, we found our framework to be insensitive to the dimensionality, resulting in a performance similar to that of analytical solutions both in low and in high dimensions. Finally, we used our framework to address the difficult and high-dimensional problem of inferring demography and selection jointly from genetic data. Specifically, and through simulations and an application to experimental data, we show that our framework enables the accurate joint estimation of the effective population size, the distribution of fitness effects of segregating mutations, and locus-specific selection coefficients from allele frequency time-series data.

More generally, we envision that any hierarchical model with genome-wide and locus-specific parameters would be well suited for application of ABC-PaSS. Such models may include hyperparameters like genome-wide mutation and recombination rates or parameters regarding the demographic history, along with locus-specific parameters that allow for between-locus variation, for instance in the intensity of selection, mutation, recombination, or migration rates. Among these, the prospect of jointly inferring selection and demographic history even from data of a single time point is particularly relevant, since it allows for the relaxation of a frequently used yet unrealistic assumption that neutral loci can be identified *a priori*. In addition, such a joint estimation allows for hierarchical parameters to aggregate information across individual loci to increase estimation power, for instance for the inference of locus-specific selection coefficients by also jointly inferring parameters of the DFE, as we did here.

Acknowledgments

We thank Pablo Duchen and the Laurent Excoffier group for comments and discussion on this work. We also thank two anonymous reviewers whose constructive criticism and comments helped to improve this work significantly. This study was supported by Swiss National Foundation grant 31003A_149920 (to D.W.).

Literature Cited

Adrion, J. R., A. Kousathanas, M. Pascual, H. J. Burrack, N. M. Haddad *et al.*, 2014 Drosophila suzukii: the genetic footprint of a recent, worldwide invasion. Mol. Biol. Evol. 31: 3148–3163.

Aeschbacher, S., M. A. Beaumont, and A. Futschik, 2012 A novel approach for choosing summary statistics in approximate Bayesian computation. Genetics 192: 1027–1047.

Aeschbacher, S., A. Futschik, and M. A. Beaumont, 2013 Approximate Bayesian computation for modular inference problems with many parameters: the example of migration rates. Mol. Ecol. 22: 987–1002.

Bank, C., G. B. Ewing, A. Ferrer-Admettla, M. Foll, and J. D. Jensen, 2014 Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 30: 540–546.

Barthelmé, S., and N. Chopin, 2014 Expectation propagation for likelihood-free inference. J. Am. Stat. Assoc. 109: 315–333.

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

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

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin, and C. P. Robert, 2009 Adaptive approximate Bayesian computation. Biometrika 96: 983–990.

Beisel, C. J., D. R. Rokyta, H. A. Wichman, and P. Joyce, 2007 Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics 176: 2441–2449.

Bilodeau, M., and D. Brenner, 2008 *Theory of Multivariate Statistics*. Springer Science & Business Media. New York, NY.

Blum, M. G. B., 2010 Approximate Bayesian computation: a nonparametric perspective. J. Am. Stat. Assoc. 105: 1178–1187.

Blum, M. G. B., M. A. Nunes, D. Prangle, and S. A. Sisson, 2013 A comparative review of dimension reduction methods in approximate Bayesian computation. Stat. Sci. 28: 189–208.

Bollback, J. P., T. L. York, and R. Nielsen, 2008 Estimation of 2nes from temporal allele frequency data. Genetics 179: 497–502.

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc. B 26: 211–252.

Brown, P. M. J., C. E. Thomas, E. Lombaert, D. L. Jeffries, A. Estoup *et al.*, 2011 The global spread of Harmonia axyridis (Coleoptera: Coccinellidae): distribution, dispersal and routes of invasion. BioControl 56: 623–641.

Casella, G., and R. L. Berger, 2002 *Statistical Inference*, Vol. 2. Duxbury Press, Pacific Grove, CA.

Chu, J.-H., D. Wegmann, C.-F. Yeh, R.-C. Lin, X.-J. Yang *et al.*, 2013 Inferring the geographic mode of speciation by contrasting autosomal and sex-linked genetic diversity. Mol. Biol. Evol. 30: 2519–2530.

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin *et al.*, 2008 Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24: 2713–2719.

Crisci, J. L., Y.-P. Poh, A. Bean, A. Simkin, and J. D. Jensen, 2012 Recent progress in polymorphism-based population genetic inference. J. Hered. 103: 287–296.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. Franccois, 2010 Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. 25: 410–418.

Durrett, R., 2008 *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media. New York, NY.

Dussex, N., D. Wegmann, and B. Robertson, 2014 Postglacial expansion and not human influence best explains the population structure in the endangered kea (Nestor notabilis). Mol. Ecol. 23: 2193–2209.

Fan, H. H., and L. S. Kubatko, 2011 Estimating species trees using approximate Bayesian computation. Mol. Phylogenet. Evol. 59: 354–363.

Fearnhead, P., and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc. Ser. B Stat. Methodol. 74: 419–474.

Foll, M., Y.-P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank *et al.*, 2014 Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 10: e1004185.

Foll, M., H. Shim, and J. D. Jensen, 2015 WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol. Ecol. Resour. 15: 87–98.

Hernandez, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786–2787.

Jabot, F., and J. Chave, 2009 Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. Ecol. Lett. 12: 239–248.

Jensen, J. D., K. R. Thornton, and P. Andolfatto, 2008 An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. PLoS Genet. 4: e1000198.

Leuenberger, C., and D. Wegmann, 2010 Bayesian computation and model selection without likelihoods. Genetics 184: 243–252.

Li, J., D. J. Nott, Y. Fan, and S. A. Sisson, 2015 Extending approximate Bayesian computation methods to high dimensions via Gaussian copula. arXiv:1504.04093.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating allele age and selection coefficient from timeserial data. Genetics 192: 599–607.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 100: 15324–15328.

Martin, G., and T. Lenormand, 2008 The distribution of beneficial and fixed mutation fitness effects close to an optimum. Genetics 179: 907–916.

Messer, P. W., 2013 SLiM: simulating evolution with selection and linkage. Genetics 194: 1037–1039.

Mevik, B., and R. Wehrens, 2007 The PLS package: principal component and partial least squares regression in R. J. Stat. Softw. 18: 1–24.

Nott, D. J., Y. Fan, L. Marshall, and S. A. Sisson, 2012 Approximate Bayesian computation and Bayes' linear analysis: toward highdimensional ABC. J. Comput. Graph. Stat. 23: 65–86.

Ratmann, O., O. Jørgensen, T. Hinkley, M. Stumpf, S. Richardson *et al.*, 2007 Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. PLoS Comput. Biol. 3: e230.

Ray, N., M. Currat, M. Foll, and L. Excoffier, 2010 SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26: 2993–2994.

Schafer, C. M., and P. E. Freeman, 2012 Likelihood-free inference in cosmology: potential for the estimation of luminosity functions, pp. 3–19 in *Statistical Challenges in Modern Astronomy V* (Lecture Notes in Statistics no. 902), edited by E. D. Feigelson and G. J. Babu. Springer-Verlag, New York.

Sisson, S. A., Y. Fan, and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA 104: 1760–1765.

Veeramah, K. R., D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins *et al.*, 2011 An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal re-sequencing data. Mol. Biol. Evol. 29: 617–630.

Wegmann, D., and L. Excoffier, 2010 Bayesian inference of the demographic history of chimpanzees. Mol. Biol. Evol. 27: 1425–1435.

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

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier, 2010 ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11: 116.

Communicating editor: M. A. Beaumont

Appendix

Proof for Theorem 1. The transition kernel $\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ associated with the Markov chain is zero if $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ differ in more than one component. If $\boldsymbol{\theta}_{-i} = \boldsymbol{\theta}'_{-i}$ for some index i, then we have

$$\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}') = p_i \rho_i(\boldsymbol{\theta}, \boldsymbol{\theta}') + (1 - r(\boldsymbol{\theta}))\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}'), \tag{A1}$$

where $\rho_i(\boldsymbol{\theta}, \boldsymbol{\theta}') = q_i(\boldsymbol{\theta}'|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}')h(\boldsymbol{\theta}, \boldsymbol{\theta}')$, $\delta_{\boldsymbol{\theta}}$ is the Dirac mass in $\boldsymbol{\theta}$, and

$$r(\boldsymbol{\theta}) = \sum_{i=1}^{n} p_i \int \rho_i\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) d\boldsymbol{\theta}'_i g.$$

We may assume without loss of generality that

$$\frac{\pi(\boldsymbol{\theta}')q_i(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q_i(\boldsymbol{\theta}'|\boldsymbol{\theta})} \leq 1.$$

From (1) we conclude

$$\mathbb{P}(\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta})g_i(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-i}).$$

Setting

$$c := \left(\int \mathbb{P}(\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right)^{-1}$$

and keeping in mind that $\boldsymbol{\theta}_{-i} = \boldsymbol{\theta}'_{-i}$ and $h(\boldsymbol{\theta}', \boldsymbol{\theta}) = 1$, we get

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})\rho_i\left(\boldsymbol{\theta}, |\boldsymbol{\theta}'\right) &= \pi(\boldsymbol{\theta}|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})q_i\left(\boldsymbol{\theta}'|\boldsymbol{\theta}\right)\mathbb{P}\left(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}'\right)h\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) \\
&= c\ \mathbb{P}(\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})q_i\left(\boldsymbol{\theta}'|\boldsymbol{\theta}\right)\mathbb{P}\left(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}'\right)\frac{\pi(\boldsymbol{\theta}')q_i(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q_i(\boldsymbol{\theta}'|\boldsymbol{\theta})} \\
&= c\ \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta})g_i(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}_{-i})\mathbb{P}\left(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}'\right)\pi\left(\boldsymbol{\theta}'\right)q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right) \\
&= c\ \mathbb{P}\left(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta}'\right)g_i\left(\boldsymbol{s}_{\text{obs}}, \boldsymbol{\theta}'_{-i}\right)\mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta})\pi\left(\boldsymbol{\theta}'\right)q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right) \\
&= c\ \mathbb{P}\left(\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}|\boldsymbol{\theta}'\right)\mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_{i,\text{obs}}|\boldsymbol{\theta})\pi\left(\boldsymbol{\theta}'\right)q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right)h\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right) \\
&= \pi\left(\boldsymbol{\theta}'|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}}\right)\rho_i\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right).
\end{aligned}$$

From this and Equation A1 it follows readily that the transition kernel $\mathcal{K}(\cdot, \cdot)$ satisfies the detailed balanced equation

$$\pi(\boldsymbol{\theta}|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}'|\boldsymbol{s} = \boldsymbol{s}_{\text{obs}})\mathcal{K}(\boldsymbol{\theta}', \boldsymbol{\theta})$$

of the Metropolis–Hastings chain.

□

Suppose that, given the parameters $\boldsymbol{\theta}$, the distribution of the statistics vector $\boldsymbol{s}$ is multivariate normal according to the GLM

$$\boldsymbol{s} = \boldsymbol{c} + \boldsymbol{C}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_s)$ and for any $m \times n$ matrix $\boldsymbol{C}$. If the prior distribution of the parameter vector is $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_\theta)$, then the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{s}_{\text{obs}}$ is

$$\boldsymbol{\theta}|\boldsymbol{s}_{\text{obs}} \sim \mathcal{N}(\boldsymbol{D}\boldsymbol{d}, \boldsymbol{D}) \tag{A2}$$

with $\boldsymbol{D} = (\boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C} + \boldsymbol{\Sigma}_\theta^{-1})^{-1}$ and $\boldsymbol{d} = \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{s}_{\text{obs}} - \boldsymbol{c}) + \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}_0$ (see, *e.g.*, Leuenberger and Wegmann 2010). We have the following:

Theorem 2. *Let $\mathbf{c}_i$ be the ith column of $\boldsymbol{C}$ and $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}_s^{-1}\mathbf{c}_i$. Moreover, let*

$$\tau_i = \tau_i(\mathbf{s}) = \boldsymbol{\beta}'_i\mathbf{s}.$$

Then τ_i is sufficient for the parameter θ_i and the collection of statistics

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)'$$

yields the same posterior (A2) *as* $\boldsymbol{s}$.

In practice, the design matrix $\boldsymbol{C}$ is unknown. We can perform an initial set of simulations from which we can infer that

$$\text{Cov}(\boldsymbol{s}, \theta_i) = \text{Var}(\theta_i)\mathbf{c}_i.$$

A reasonable estimator for the sufficient statistic τ_i is then $\hat{\tau}_i = \widehat{\beta}'_i s$ with

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\Sigma}}_s^{-1}\hat{\boldsymbol{\Sigma}}_{s\theta_i}, \tag{A3}$$

where $\hat{\boldsymbol{\Sigma}}_s$ and $\hat{\boldsymbol{\Sigma}}_{s\theta_i}$ for $i = 1, \ldots, n$ are the covariances estimated, for instance, using ordinary least squares.

Proof for Theorem 2. It is easy to check that the mean of τ_i is $\mu_i = \mathbf{t}'_i(\boldsymbol{C\theta} + \mathbf{c})$ and its variance is $\sigma_i^2 = \mathbf{t}_i{}'\boldsymbol{\Sigma}_s\mathbf{t}_i{}'$. The covariance between $\boldsymbol{s}$ and τ is given by

$$\begin{aligned} \boldsymbol{\Sigma}_{s\tau} &= \mathbb{E}((\boldsymbol{s} - \boldsymbol{C\theta} - \mathbf{c})(\tau_i - \mu_i)) \\ &= \mathbb{E}(\boldsymbol{\epsilon\epsilon}'\boldsymbol{t}_i) = \boldsymbol{\Sigma}_s\boldsymbol{t}_i \end{aligned}$$

Consider the conditional multinormal distribution $\boldsymbol{s}|\tau_i$. Using the well-known formula for the variance and the mean of a conditional multivariate normal (see, *e.g.*, Bilodeau and Brenner 2008), we get that the covariance of $\boldsymbol{s}|\tau_i$ is given by

$$\boldsymbol{\Sigma}_{s|\tau} = \boldsymbol{\Sigma}_s - \sigma_i^{-2}\boldsymbol{\Sigma}_{s\tau}\boldsymbol{\Sigma}'_{s\tau}$$

and thus is independent of $\boldsymbol{\theta}$. The mean of $\boldsymbol{s}|\tau_i$ is

$$\boldsymbol{\mu}_{s|\tau} = \boldsymbol{C\theta} + \mathbf{c} + \sigma_i^{-2}\boldsymbol{\Sigma}_{s\tau}\boldsymbol{t}_i{}'(\boldsymbol{s} - \boldsymbol{C\theta} - \boldsymbol{c}).$$

The part of this expression depending on θ_i is

$$\left(\boldsymbol{I} - \frac{\boldsymbol{\Sigma}_s\boldsymbol{t}_i\boldsymbol{t}_i{}'}{\boldsymbol{t}_i{}'\boldsymbol{\Sigma}_s\boldsymbol{t}_i}\right)\mathbf{c}_i\theta_i.$$

Inserting $\boldsymbol{t}_i = \boldsymbol{\Sigma}_s^{-1}\boldsymbol{c}_i$ we obtain

$$\left(c_i - \frac{\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_s^{-1}\boldsymbol{c}_i\boldsymbol{c}_i{}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{c}_i}{\boldsymbol{c}_i{}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_s^{-1}\boldsymbol{c}_i}\right)\theta_i = (\mathbf{c}_i - \mathbf{c}_i)\theta_i = \mathbf{0}.$$

Thus the distribution of $\boldsymbol{s}|\tau_i$ is independent of θ_i and hence τ_i is sufficient for θ_i.

To prove the second part of *Theorem 2*, we observe that $\boldsymbol{\tau}$ is given by the linear model

$$\boldsymbol{\tau} = \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{s} = \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C\theta} + \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\mathbf{c} + \boldsymbol{\eta}$$

with $\boldsymbol{\eta} = \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\epsilon}$. Using $\text{Cov}(\boldsymbol{\eta}) = \boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C}$ we get for the posterior variance

$$\left(\boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\left(\boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} = \left(\boldsymbol{C}'\boldsymbol{\Sigma}_s^{-1}\boldsymbol{C} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} = \boldsymbol{D}.$$

Similarly we see that the posterior mean is $\boldsymbol{Dd}$.

□

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187567/-/DC1

Likelihood-Free Inference in High-Dimensional Models

Athanasios Kousathanas, Christoph Leuenberger, Jonas Helfer, Mathieu Quinodoz, Matthieu Foll, and Daniel Wegmann

DOI: 10.1534/genetics.116.187567

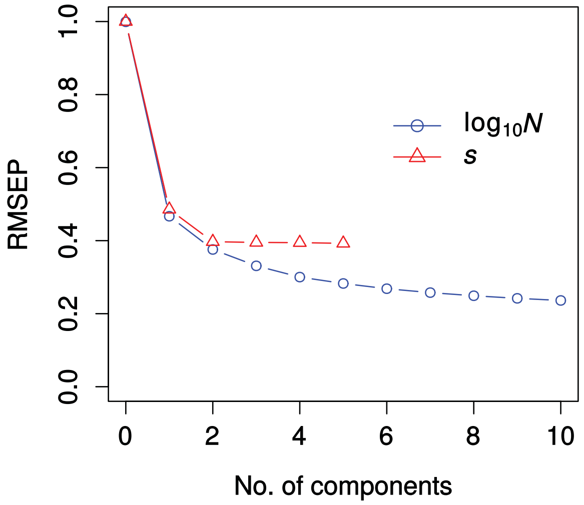

Figure S1: **Prediction error for partial least squares analysis (PLS).** The root mean squared prediction error (RMSEP) is shown for parameters N_e and s as a function of increasing number of PLS components. We performed PLS using Wright-Fisher simulations of 100 loci where we assumed uniform priors for N_e and s ($U[0.5, 4.5]$ and $U[0, 1]$, respectively).

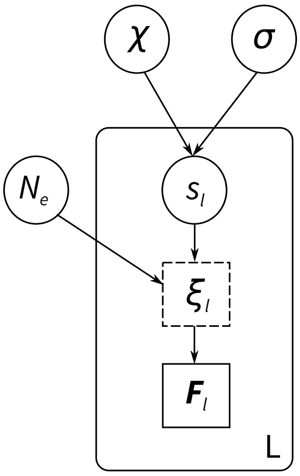

Figure S2: **Directed acyclic graph describing the Wright-Fisher model examined in this study.** Solid circles represent parameters to be estimated. The dashed square represents the full data, which is summerized here by a vector of statistics $\boldsymbol{F}_l$, indicated by a solid square. Nodes contained in the plate are repeated for each locus $l \in \{1, \ldots, L\}$ times.

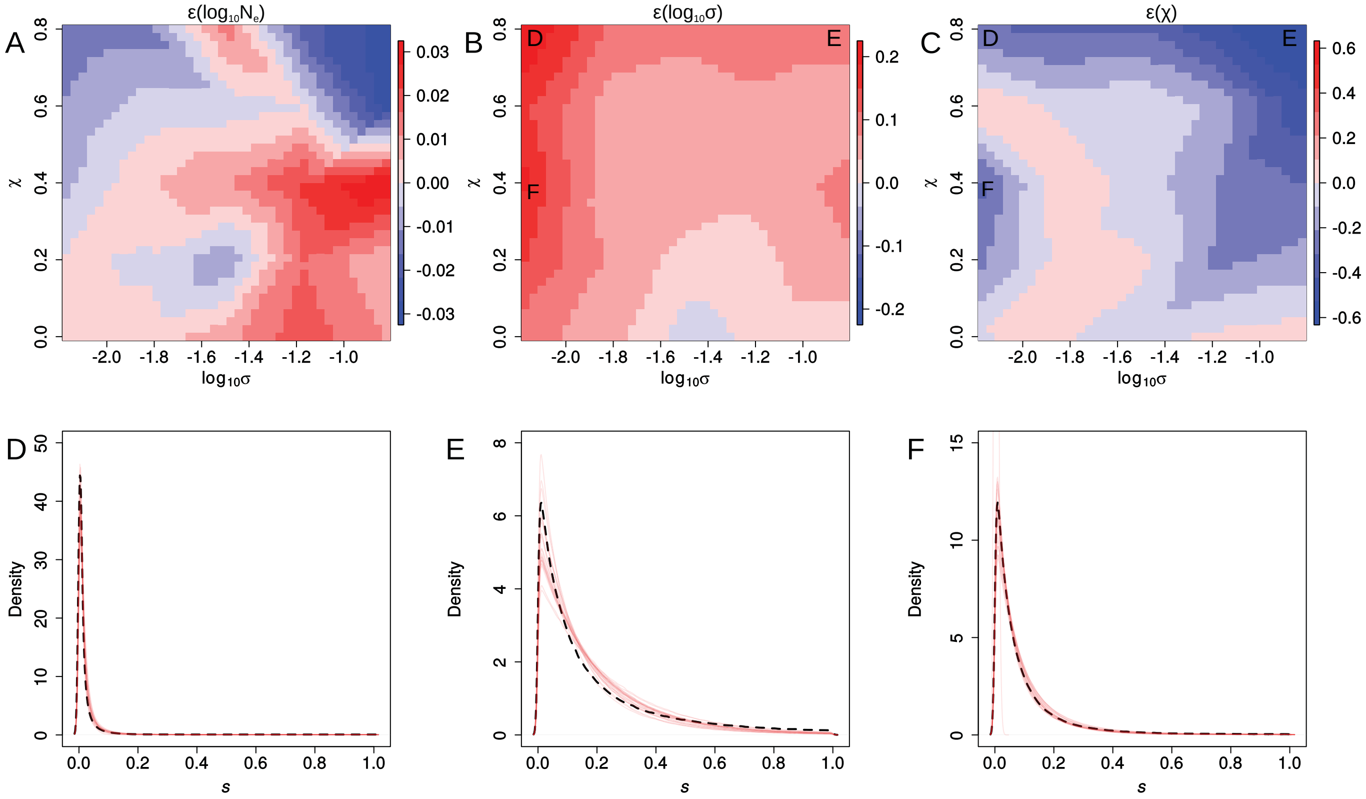

Figure S3: **Accuracy in estimating N_e and DFE parameters σ and χ jointly.** (A,B,C) Sets of simulations of 100 loci were conducted for combinations of parameters σ and χ over a grid from their prior range and we evaluated the median approximation error (ϵ=estimate-true) over 25 replicates. Color gradients indicate the extent of overestimation (red) or underestimation (blue) of each parameter. These results suggest very high accuracy when estimating N_e with maximum $\epsilon \approx 0.04$ or 1% of the prior range and rather low for σ (about 10% of the prior range). In contrast, ϵ is rather large for χ, spanning up to 75% of the prior range. This is due to several combinations of χ and σ leading to very similar shapes of the truncated GPD. This is illustrated in panels D, E anf F, where we show the true (dashed black line) versus estimated (red) DFE obtained for 50 replicates using parameter combinations of χ and σ as indicated in panels B and C.

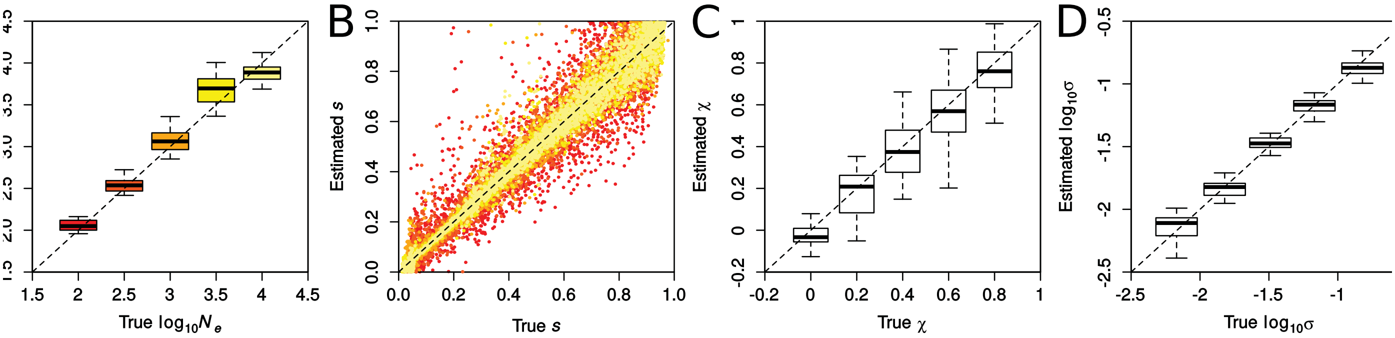

Figure S4: **Accuracy in inferring demographic and selection parameters using the PLS 5/2 set of statistics.** Results were obtained with `ABC-PaSS` using five and two PLS components for N_e and each s, respectively (PLS 5/2). Shown are the true versus estimated posterior medians for parameters N_e (A), s per locus (B), χ and σ of the Generalized Pareto distribution (C and D, respectively). Boxplots summarize results from 25 replicate simulations, each with 100 loci. Uniform priors over the whole ranges shown were used. (A, B): N_e assumed in the simulations is represented as a color gradient of red (low N_e) to yellow (high N_e). (C,D): Parameters μ and N_e were fixed to 0 and 10^3, respectively, $log_{10}\sigma$ was fixed to -1 (C) and χ was fixed to 0.5 (D).

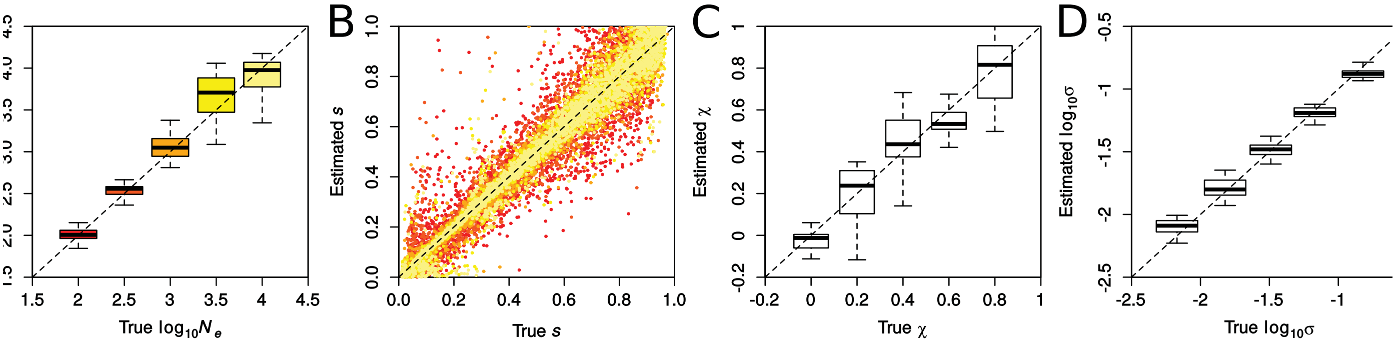

Figure S5: **Accuracy in inferring demographic and selection parameters using the PLS $3/1$ set of statistics.** Results were obtained with ABC-PaSS using three and one PLS components for N_e and each s, respectively (PLS 5/2). Shown are the true versus estimated posterior medians for parameters N_e (A), s per locus (B), χ and σ of the Generalized Pareto distribution (C and D, respectively). Boxplots summarize results from 25 replicate simulations, each with 100 loci. Uniform priors over the whole ranges shown were used. (A, B): N_e assumed in the simulations is represented as a color gradient of red (low N_e) to yellow (high N_e). (C,D): Parameters μ and N_e were fixed to 0 and 10^3, respectively, $log_{10}\sigma$ was fixed to -1 (C) and χ was fixed to 0.5 (D).

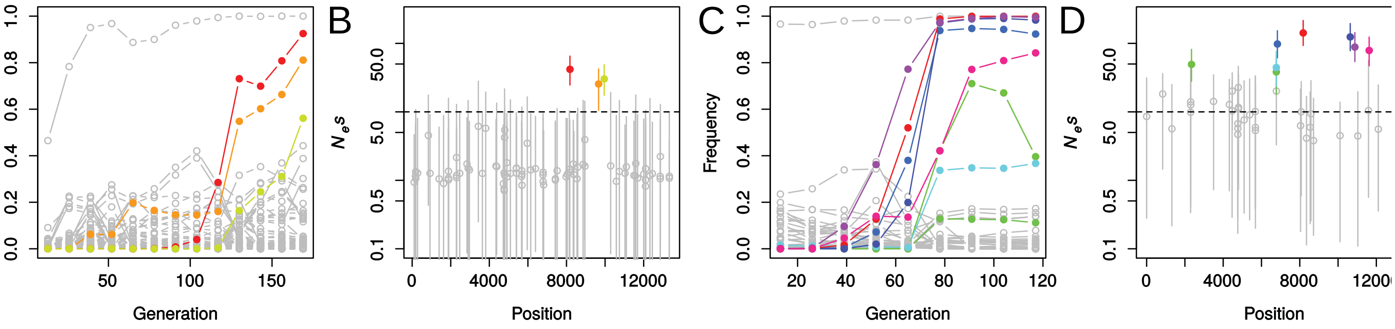

Figure S6: **Allele trajectories (A, C) and posterior estimates for $N_e s$ (B, D) for control (A, B) and drug-treated (C, D) Influenza.** Non-significant loci are colored grey and significant loci are colored with a unique color for each locus.

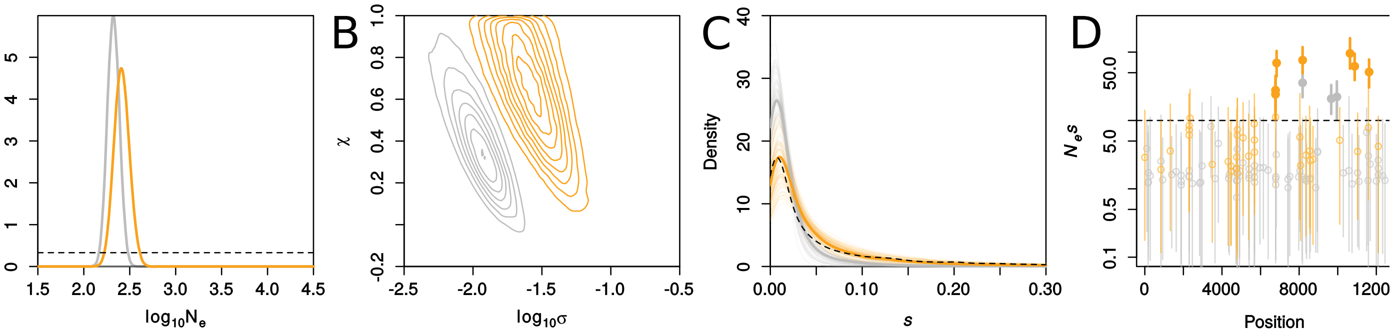

Figure S7: Inferred demography and selection for experimental evolution of Infuenza using the PLS 3/1 set of statistics. We show results for the no-drug (control) and drug-treated Influenza in grey and orange, respectively. Shown are the posterior distributions for $log_{10}N_e$ (A) and $log_{10}\sigma$ and χ (B). In panel C, we plotted the modal distribution of fitness effects (DFE) with thick lines by integrating over the posterior of its parameters. The thin lines represent the DFEs obtained by drawing 100 samples from the posterior of σ and χ. Dashed lines in panels A and C correspond to the prior distributions. In panel D, the posterior estimates for Nes per locus versus the position of the loci in the genome are shown. Open circles indicate non-significant loci whereas closed, thick circles indicate significant loci ($P(N_e s > 10) > 0.95$, dashed line).

Simulation set	Parameter	Method	RMSE	Pearsons R^2
Set 1	s	LC 1/1	0.0700	0.970
		PLS 3/1	0.0809	0.960
		PLS 5/2	0.0743	0.966
	$\log_{10}(N_e)$	LC 1/1	0.178	0.969
		PLS 3/1	0.199	0.962
		PLS 5/2	0.171	0.972
Set 2	s	LC 1/1	0.0191	0.954
		PLS 3/1	0.0390	0.957
		PLS 5/2	0.0195	0.953
	$\log_{10}(N_e)$	LC 1/1	0.0292	-
		PLS 3/1	0.0390	-
		PLS 5/2	0.0485	-
	χ	LC 1/1	0.126	0.911
		PLS 3/1	0.127	0.910
		PLS 5/2	0.140	0.895
Set 3	s	LC 1/1	0.0226	0.987
		PLS 3/1	0.0228	0.986
		PLS 5/2	0.0227	0.986
	$\log_{10}(N_e)$	LC 1/1	0.0427	-
		PLS 3/1	0.189	-
		PLS 5/2	0.374	-
	log10(σ)	LC 1/1	0.0783	0.988
		PLS 3/1	0.0776	0.989
		PLS 5/2	0.0844	0.984
Set 4	s	LC 1/1	0.0258	0.982
		PLS 3/1	0.0259	0.981
		PLS 5/2	0.0254	0.981
	$\log_{10}(N_e)$	LC 1/1	0.0577	-
		PLS 3/1	0.0889	-
		PLS 5/2	0.351	-
	$\log_{10}(\sigma)$	LC 1/1	0.0191	0.961
		PLS 3/1	0.0180	0.968
		PLS 5/2	0.0167	0.968
	χ	LC 1/1	0.306	0.667
		PLS 3/1	0.306	0.659
		PLS 5/2	0.291	0.666

Table S1: **Performance of ABC-PaSS coupled with different dimension reduction techniques in Wright-Fisher simulations.** We computed the root mean square error (RMSE) and Pearson's correlation (R^2) between true and estimated parameter values using three different dimension reduction strategies: a single linear combination per parameter calculated according to Theorem 2 (LC 1/1), three and one PLS components for parameters $log_{10}N_e$ and s, respectively (PLS 3/1), and five and two PLS components for parameters $log_{10}N_e$ and s, respectively (PLS 5/2). Results shown are for four sets of 25 replicate simulations: Set 1 assumed uniform priors for N_e and s ($U[0.5, 4.5]$ and $U[0, 1]$, respectively), Sets 2-4 assumed a generalised pareto distribution for s with hyperparamers σ and χ. For Set 2 we varied χ ($U[-0.2, 1]$) and kept σ fixed ($= 0.01$), for set 3 we varied $log_{10}\sigma$ ($U[-2.5, -0.5]$) and kept χ fixed ($= 0.5$) and for Set 4 we varied both σ and χ. For Sets 2-4 we performed simulations only for $log_{10}N_e = 3$, thus R^2 is not calculable for these. The dimension reduction strategy with the smallest RMSE for each parameter per set is highlighted in grey.