Likelihood-Free Inference in High-Dimensional Models

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

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 calculated from the observed data is sufficient even for this parameter. This increases acceptance rates dramatically, rendering this approach 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 even for models of very high dimensionality. We further derive that for linear models, a one-dimensional combination of statistics per parameter is sufficient for this parameter. This increases acceptance rates dramatically, rendering this approach 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 even for models of very high dimensionality.

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.

Supplemental material is available online

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

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

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

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

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

However, for models with $m \gg 1$ the condition $s = s_{\text{obs}}$ might be too restrictive and require a prohibitively large simulation effort. Therefore, an approximation step can be employed by relaxing the condition $s = s_{\text{obs}}$ to $\|s - s_{\text{obs}}\| \leq \delta$, where
\( \|x - y\| \) is a distance metric of choice between \( x \) and \( y \) and \( \delta \) is a chosen distance (tolerance) below which simulations are accepted. The posterior \( \pi(\theta|s) \) is thus approximated by

\[
\hat{\pi}(\theta|s) = \frac{\mathbb{P}(\|s - s_{\text{obs}}\| \leq \delta \| \theta \) \pi(\theta) \mathbb{I}}{\int \mathbb{P}(\|s - s_{\text{obs}}\| \leq \delta \| \theta \) \pi(\theta) \mathbb{I} \mathbb{d}\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 \( \theta \) propose to move to \( \theta' \) according to the transition kernel \( q(\theta'|\theta) \).
2. Simulate \( D \) using model \( M \) with \( \theta' \) and calculate summary statistics \( s \) for \( D \).
3. If \( \|s - s_{\text{obs}}\| \leq \delta \), go to step 4; otherwise go to step 1.
4. Calculate the Metropolis–Hastings ratio

\[ h(h, \theta', \theta) = \min(1, \frac{\pi(\theta')q(\theta'|\theta)}{\pi(\theta)q(\theta'|\theta)}) \]

5. Accept \( \theta' \) with probability \( h \); otherwise stay at \( \theta \). 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 \( \delta \). In such situations, the condition \( \|s - s_{\text{obs}}\| \leq \delta \) 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 \( T_i = T_i(s) \) as an \( m_i \)-dimensional function of \( s \). We call \( T_i \) sufficient for the parameter \( \theta_i \) if the conditional distribution of \( s \) given \( T_i \) does not depend on \( \theta_i \). More precisely, let \( t_{obs} = T_i(s_{obs}) \). Then

\[
\mathbb{P}(s = s_{obs}|T_i = t_{obs}, \theta) = \frac{\mathbb{P}(s = s_{obs}, T_i = t_{obs}|\theta)}{\mathbb{P}(T_i = t_{obs}|\theta)}
\]

\[
= \frac{\mathbb{P}(s = s_{obs}|\theta)}{\mathbb{P}(T_i = t_{obs}|\theta)} = : g_i(s_{obs} \theta_{-i}),
\]

where \( \theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_0) \) is \( \theta \) with the \( i \)th 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 \( s \) has the form

\[
f(s|\theta) = h(s) \exp \left[ \sum_{k=1}^{K} \eta_k(\theta)T_k(s) - A(\theta) \right].
\]
For a given parameter $\theta_i$, the vector $T_i(s)$ consisting of only those $T_{ij}(s)$ for which the respective natural parameter function $\eta_j(\theta)$ depends on $\theta_i$ is a sufficient statistic for $\theta_i$ in the sense of our definition. Some concrete examples of this type are studied below.

If sufficient statistics $T_i$ can be found for each parameter $\theta_i$ and their dimension $m_i$ is substantially smaller than the dimension $m$ of $s$, then the ABC-MCMC algorithm can be greatly improved with the following algorithm that we denote ABC with parameter-specific statistics (ABC-PaSS) henceforth.

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

1. Choose an index $i = 1, \ldots, n$ according to a probability distribution $(p_1, \ldots, p_n)$ with $\sum p_i = 1$ and all $p_i > 0$.
2. At $\theta = \theta^{(i)}$ propose $\theta'$ according to the transition kernel $q_i(\theta' | \theta)$ where $\theta'$ differs from $\theta$ only in the $i$th component:

   $\theta' = (\theta_1, \ldots, \theta_{i-1}, \theta'_i, \theta_{i+1}, \ldots, \theta_n)$.

3. Simulate $D$ using model $M$ with $\theta'$ and calculate summary statistics $s$ for $D$. Calculate $t_i = T_i(s)$ and $t_{i, \text{obs}} = T_i(s_{\text{obs}})$.
4. Let $\delta_i$ be the tolerance for parameter $\theta_i$. If $||t_i - t_{i, \text{obs}}||_i \leq \delta_i$, go to step 5; otherwise go to step 1.
5. Calculate the Metropolis–Hastings ratio

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

6. Accept $\theta'$ with probability $h$; otherwise stay at $\theta$.
7. Increase $t$ by one, save a new parameter value $\theta^{(i)} = \theta$, and continue at step 1.

Convergence of the MCMC chain is guaranteed by the following:

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

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 $\pi(\theta | s = s_{\text{obs}})$. Because of the continuity of the summary statistics $s$ and the sufficient statistics $T_i$, 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 $\theta^{(1)}$ and tolerance $\delta_i$ 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 $\bar{x}$ and sample variance $S^2$ of samples of size $n$ as statistics. Recall that for noninformative priors the posterior distribution for $\mu$ is $N(\bar{x}, S^2/n)$ and the posterior distribution for $\sigma^2$ is $\chi^2$ distributed with $n - 1$ d.f. As $\mu$ and $\sigma^2$ are independent, we get the posterior density

$$\pi(\mu, \sigma^2) = \phi_{\bar{x}, S^2/n}(\mu) \cdot \frac{n-1}{S^2} \Gamma(n/2, 1/\sigma^2).$$

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 $\mu$ and $\sigma^2$. 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 $s$ being a linear function of $n = m$ parameters $\theta$, $s = C\theta + \epsilon$, $\epsilon \sim N(0, I)$, where $C$ is a square design matrix and the vector of errors $\epsilon$ is multivariate normal. Under noninformative priors for the parameters $\theta$, their posterior distribution is multivariate normal

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

We set up the design matrices $C$ 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 $C = B \cdot \det(B^T B)^{-1/2n}$, where

$\begin{align*}
C &= \begin{pmatrix}
1 & \cdots & 1
\end{pmatrix} \\
\epsilon &= \begin{pmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_m
\end{pmatrix} \\
N(0, I) &= \text{Normal distribution}
\end{align*}$
The normalization factor in the definition of \( C \) 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 of \( n \). 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 \( \theta = 0 \) and the priors are uniform \( U[-100, 100] \) for all parameters, which are effectively non-informative. We started the MCMC chains at a normal deviate \( \mathcal{N}(\theta, 0.01I) \), i.e., around the true values of \( \theta \). 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 \( \xi \) of observed allele trajectories (sample allele frequencies) over \( l = 1, \ldots, L \) 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 \( s_i \), respectively, under the classic Wright–Fisher model with allelic fitnesses 1 and 1 + \( s_i \). We further assume the locus-specific selection coefficients \( s_i \) follow a DFE parameterized as a generalized Pareto distribution (GPD) with mean \( \mu = 0 \), shape \( \chi \), and scale \( \sigma \). Our goal is thus to estimate the joint posterior distribution

\[
\pi(N_e, s_1, \ldots, s_L, \chi, \sigma | \xi) \propto \prod_{l=1}^{L} \left[ \mathcal{P}(\xi_l | N_e, s_l) \pi(s_l | \chi, \sigma) \right] \pi(N_e) \pi(\chi) \pi(\sigma).
\]

To apply our ABC-PaSS framework to this problem, we approximate the likelihood term \( \mathcal{P}(\xi_l | N_e, s_l) \) numerically with simulations, while updating the hyperparameters \( \chi \) and \( \sigma \) analytically.

**Summary statistics:** To summarize the data \( \xi \), 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

\[
Fs' = \frac{(x-y)^2}{z(1-z)}
\]

where \( x \) and \( y \) are the minor allele frequencies separated by \( t \) generations, \( z = (x + y) / 2 \), and \( h \) 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 \( Fs'd'_{i} \) 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 \( F_l = \{Fs'_{i}, Fs'd'_{i}, Fs'_{i}^2, Fs'd'_{i}^2, Fs'_{i} \times Fs'd'_{i}\} \)

We next calculated parameter-specific linear combinations for \( N_e \) and locus-specific \( s_i \) 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 \( F_l \) 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 \( F \) for each parameter \( N_e \) and \( s \). This resulted in \( \tau(F_l | F) = \beta F_l \) and \( \tau(N, F) = \beta N F_l \). To combine information across loci when updating \( N_e \), we then calculated

\[
\tau_{N_e}(F) = \sum_{l=1}^{L} \beta F_l F_l,
\]

where \( F = \{F_1, \ldots, F_L\} \). In summary, we used the ABC approximation

\[
P(\xi_l | N_e, s_l) \approx P\left( \|\tau_s(F_l) - \tau_{\text{true}}(F_{\text{true}})\| < \delta_{s}, \|\tau_{N_e}(F) - \tau_{N_e}(F_{\text{true}})\| < \delta_{N_e} | N_e, s_l \right).
\]

**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 \( \sim 13 \) 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.unmassmed.edu/influenza/](http://bib.unmassmed.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
We simulated 13 generations between time points and a house C++ code implemented as a module of ABCtoolbox. time points during which drug was administered. of the data of the drug-treated experiment to the last nine were 86 and 42 such loci for the control and drug-treated estimated posterior distributions for parameter-speci

The authors state that all data necessary for con

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 $\mu$ (A) and $\sigma^2$ (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 $\mu$ (C) and $\sigma^2$ (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.

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 ($\mu$) and variance ($\sigma^2$). Given a sample of size $n$, the sample mean ($\bar{x}$) is a sufficient statistic for $\mu$, while both $\bar{x}$ and the sample variance ($S^2$) are sufficient for $\sigma^2$ (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 $\mu$ and both $\bar{x}$ and $S^2$ when updating $\sigma^2$.

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 $\delta_i$ 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.

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 log10 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 $\delta_i$ 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.
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 \( s \) while for ABC-PaSS, we employed Theorem 2 and used a single linear combination of statistics \( \tau_i \) per parameter \( \theta_i \). 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
empathically 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 in the exponential family, 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 summary 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 $\mu$, 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 $\mu$ and $N_e$ chosen using ordinary least squares, as suggested by Fearnhead and Prangle (2012) (LC 1/1); (2) two PLS components for $\mu$ 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 $\mu$ 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 $\mu$ 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 $\mu$ were on average unbiased and accuracy was, as expected, higher for larger $N_e$ (Figure 3B). Note that since the prior on $\mu$ 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 $N_e \log_2 x > 10$). 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 $\mu$ is realistically described by a truncated GPD with location $\mu = 0$ and parameters shape $\sigma$ and scale $\chi$ (Figure S2).

We first evaluated the accuracy of estimating $\chi$ and $\sigma$ 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 $\chi$ and $\sigma$ is very similar, these parameters are not always identifiable. This renders the

![Figure 3](image-url)  
**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 $\mu$. Shown are the true vs. estimated posterior medians for parameters $N_e$ (A), $\mu$ per locus (B), and $\chi$ and $\sigma$ 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 $\mu$ and $N_e$ were fixed to 0 and $10^3$, respectively; $\log_2 \chi$ was fixed to $-1$ (C), and $\chi$ was fixed to 0.5 (D).
and we found the effective population size on set 1 (Table S1). However, in the presence of hyperparameters, particularly when uniform priors were used on all the differences between LC 1/1 and PLS 5/2 were very subtle, set (PLS 3/1) performed worst in all comparisons, while from a previous study (Foll et al. 2014), results were more variable (simulation sets 2 – 4; Table S1). These estimates 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 influenza 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 influenza 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 \( N_e \) 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 \( \chi \) parameters for the drug-treated and for the control influenza (posterior medians 0.44 and 0.56, respectively), the \( \sigma \) 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 \( N_e \) 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


Communicating editor: M. A. Beaumont
Appendix

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

$$
\mathcal{K}(\theta, \theta') = p_i p_i(\theta, \theta') + (1 - r(\theta))\delta_\theta(0),
$$

(A1)

where $p_i(\theta, \theta') = q_i(\theta' | \theta)P(T_i = t_{i, \text{obs}}| \theta')h(\theta', \theta')$. $\delta_\theta$ is the Dirac mass in $\theta$, and

$$
r(\theta) = \sum_{i=1}^{n} p_i \int p_i(\theta, \theta') d\theta'.
$$

We may assume without loss of generality that

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

From (1) we conclude

$$
P(s = s_{\text{obs}} | \theta) = P(T_i = t_{i, \text{obs}} | \theta)^{g_i(s_{\text{obs}}, \theta_{-i})}.
$$

Setting

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

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

$$
\begin{align*}
\pi(\theta | s = s_{\text{obs}})p_i(\theta | \theta') &= \pi(\theta | s = s_{\text{obs}})q_i(\theta' | \theta)P(T_i = t_{i, \text{obs}}| \theta')h(\theta, \theta') \\
&= c P(s = s_{\text{obs}} | \theta) \pi(\theta) q_i(\theta' | \theta) P(T_i = t_{i, \text{obs}}| \theta') \frac{\pi(\theta')q_i(\theta' | \theta)}{\pi(\theta)q_i(\theta' | \theta)} \\
&= c P(T_i = t_{i, \text{obs}} | \theta)^{g_i(s_{\text{obs}}, \theta_{-i})} P(T_i = t_{i, \text{obs}}| \theta') \pi(\theta') q_i(\theta' | \theta) \\
&= c P(T_i = t_{i, \text{obs}} | \theta')^{g_i(s_{\text{obs}}, \theta'_{-i})} P(T_i = t_{i, \text{obs}}| \theta) \pi(\theta') q_i(\theta' | \theta) \\
&= c P(s = s_{\text{obs}} | \theta') P(T_i = t_{i, \text{obs}} | \theta) \pi(\theta') q_i(\theta' | \theta) h(\theta', \theta) \\
&= \pi(\theta' | s = s_{\text{obs}}) p_i(\theta', \theta).
\end{align*}
$$

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

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

of the Metropolis–Hastings chain. 

\[\square\]

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

$$
s = c + C\theta + \epsilon,
$$

where $\epsilon \sim \mathcal{N}(0, \Sigma_s)$ and for any $m \times n$ matrix $C$. If the prior distribution of the parameter vector is $\theta \sim \mathcal{N}(\theta_0, \Sigma_\theta)$, then the posterior distribution of $\theta$ given $s_{\text{obs}}$ is

$$
\theta | s_{\text{obs}} \sim \mathcal{N}(D_d, D)
$$

(A2)
with \( D = (C \Sigma_x^{-1} C + \Sigma_\theta^{-1})^{-1} \) and \( d = C \Sigma_x^{-1} (s_{\text{obs}} - c) + \Sigma_\theta^{-1} \theta_0 \) (see, e.g., Leuenberger and Wegmann 2010). We have the following:

**Theorem 2.** Let \( ci \) be the \( i \)th column of \( C \) and \( \beta_i = \Sigma_x^{-1} ci \). Moreover, let

\[
\tau_i = \tau_i(s) = \beta_i's.
\]

Then \( \tau_i \) is sufficient for the parameter \( \theta_i \) and the collection of statistics

\[
\tau = (\tau_1, \ldots, \tau_n)'
\]

yields the same posterior (A2) as \( s \).

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

\[
\text{Cov}(s, \theta_i) = \text{Var}(\theta_i)c_i.
\]

A reasonable estimator for the sufficient statistic \( \tau_i \) is then \( \hat{\tau}_i = \hat{\beta}_i's \) with

\[
\hat{\beta}_i = \tilde{\Sigma}_x^{-1} \Sigma_{\theta i i},
\]

where \( \tilde{\Sigma}_x \) and \( \Sigma_{\theta i i} \) 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 \( \tau_i \) is \( \mu_i = t'_i(C\theta + c) \) and its variance is \( \sigma^2 = t'_i(\Sigma_\theta t'_i) \). The covariance between \( s \) and \( \tau \) is given by

\[
\Sigma_{s\tau} = \Sigma_s - \sigma^2 \Sigma_\theta \Sigma'_\theta,
\]

and thus is independent of \( \theta \). The mean of \( s|\tau_i \) is

\[
\mu_{s|\tau_i} = C\theta + c + \sigma^2 \Sigma_\theta t'_i (s - C\theta - c).
\]

The part of this expression depending on \( \theta_i \) is

\[
\left( I - \frac{\Sigma_\theta t'_i}{t'_i \Sigma_\theta t_i} \right) c_i \theta_i.
\]

Inserting \( t_i = \Sigma_x^{-1} c_i \) we obtain

\[
\left( c_i - \Sigma_x^{-1} c_i \Sigma_x^{-1} c_i \right) \theta_i = (c_i - c_i) \theta_i = 0.
\]

Thus the distribution of \( s|\tau_i \) is independent of \( \theta_i \) and thus \( \tau_i \) is sufficient for \( \theta_i \).

To prove the second part of Theorem 2, we observe that \( \tau \) is given by the linear model

\[
\tau = C \Sigma_x^{-1} s = C \Sigma_x^{-1} C\theta + C \Sigma_x^{-1} c + \eta
\]

with \( \eta = C \Sigma_x^{-1} e \). Using \( \text{Cov}(\eta) = C \Sigma_x^{-1} C \) we get for the posterior variance

\[
\left( C \Sigma_x^{-1} (C \Sigma_x^{-1} C) C \Sigma_x^{-1} + \Sigma_\theta^{-1} \right)^{-1} = \left( C \Sigma_x^{-1} C + \Sigma_\theta^{-1} \right)^{-1} = D.
\]

Similarly we see that the posterior mean is \( Dd \).
Likelihood-Free Inference in High-Dimensional Models

Athanasios Kousathanas, Christoph Leuenberger, Jonas Helfer, Mathieu Quinodoz, Matthieu Foll, and Daniel Wegmann
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 summarized here by a vector of statistics \( F_l \), 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 $\sigma$ and $\chi$ jointly. (A,B,C) Sets of simulations of 100 loci were conducted for combinations of parameters $\sigma$ and $\chi$ over a grid from their prior range and we evaluated the median approximation error ($\epsilon=\text{estimate-true}$) 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 $\sigma$ (about 10% of the prior range). In contrast, $\epsilon$ is rather large for $\chi$, spanning up to 75% of the prior range. This is due to several combinations of $\chi$ and $\sigma$ leading to very similar shapes of the truncated GPD. This is illustrated in panels D, E and F, where we show the true (dashed black line) versus estimated (red) DFE obtained for 50 replicates using parameter combinations of $\chi$ and $\sigma$ 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), $\chi$ and $\sigma$ 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 $\mu$ and $N_e$ were fixed to 0 and $10^3$, respectively, $\log_{10} \sigma$ was fixed to -1 (C) and $\chi$ 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 3/2). Shown are the true versus estimated posterior medians for parameters \( N_e \) (A), \( s \) per locus (B), \( \chi \) and \( \sigma \) 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 \( \mu \) and \( N_e \) were fixed to 0 and \( 10^3 \), respectively, \( \log_{10} \sigma \) was fixed to -1 (C) and \( \chi \) was fixed to 0.5 (D).
Figure S6: Allele trajectories (A, C) and posterior estimates for $N_e$ (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 Influenza 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 $\chi$ (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 $\sigma$ and $\chi$. Dashed lines in panels A and C correspond to the prior distributions. In panel D, the posterior estimates for $N_{es}$ 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 ($P(N_{es} > 10) > 0.95$, dashed line).
<table>
<thead>
<tr>
<th>Simulation set</th>
<th>Parameter</th>
<th>Method</th>
<th>RMSE</th>
<th>Pearson’s $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>$s$</td>
<td>LC 1/1</td>
<td>0.0700</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0809</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.0743</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td>$\log_{10}(N_c)$</td>
<td>LC 1/1</td>
<td>0.178</td>
<td>0.969</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.199</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.171</td>
<td>0.972</td>
</tr>
<tr>
<td>Set 2</td>
<td>$\log_{10}(N_c)$</td>
<td>LC 1/1</td>
<td>0.0292</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0390</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.0485</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\chi$</td>
<td>LC 1/1</td>
<td>0.126</td>
<td>0.911</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.127</td>
<td>0.910</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.140</td>
<td>0.895</td>
</tr>
<tr>
<td>Set 3</td>
<td>$\log_{10}(N_c)$</td>
<td>LC 1/1</td>
<td>0.0226</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0228</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.0227</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>$\log_{10}(\sigma)$</td>
<td>LC 1/1</td>
<td>0.0783</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0776</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.0844</td>
<td>0.984</td>
</tr>
<tr>
<td>Set 4</td>
<td>$\log_{10}(\sigma)$</td>
<td>LC 1/1</td>
<td>0.0258</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0259</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.0254</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>$\chi$</td>
<td>LC 1/1</td>
<td>0.0577</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 3/1</td>
<td>0.0889</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLS 5/2</td>
<td>0.351</td>
<td>-</td>
</tr>
</tbody>
</table>

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 $\log_{10} N_c$ and $s$, respectively (PLS 3/1), and five and two PLS components for parameters $\log_{10} N_c$ and $s$, respectively (PLS 5/2). Results shown are for four sets of 25 replicate simulations: Set 1 assumed uniform priors for $N_c$ and $s$ ($U[0.5, 4.5]$ and $U[0, 1]$, respectively). Sets 2-4 assumed a generalised pareto distribution for $s$ with hyperparameters $\sigma$ and $\chi$. For Set 2 we varied $\chi$ ($U[-0.2, 1]$) and kept $\sigma$ fixed (= 0.01), for set 3 we varied $\log_{10} \sigma$ ($U[-2.5, -0.5]$) and kept $\chi$ fixed (= 0.5) and for Set 4 we varied both $\sigma$ and $\chi$. For Sets 2-4 we performed simulations only for $\log_{10} N_c = 3$, 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.