bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Joint inference of adaptive and demographic history from temporal population genomic data

Vitor A. C. Pavinato[1,2,3], Stéphane De Mita[4], Jean-Michel Marin[2], Miguel de Navascués[1,5*]

[1]CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, Université de Montpellier, Montpellier, France
[2]UMR Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France
[3]Entomology Dept., CFAES, The Ohio State University, Wooster, USA
[4]UMR Interactions Arbres-Microorganismes, INRAE, France
[5]Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden
*corresponding author: miguel.navascues@inrae.fr

October 16, 2021

Abstract

Disentangling the effects of selection and drift is a long-standing problem in population genetics. Recently, simulations shows that pervasive selection may bias the inference of demography. Ideally, models for the inference of demography and selection should account for the interaction between these two forces. With simulation-based likelihood-free methods such as Approximate Bayesian Computation, demography and selection parameters can be jointly estimated (ABC). We propose a ABC-Random Forests framework to jointly infer demographic and selection parameters from temporal population genomic data (e.g. experimental evolution, monitored populations, ancient DNA). Our framework allowed the separation of demography (census size, N) from the genetic drift (effective population size, N_{e}), and the estimation of genome-wide parameters of selection. Selection parameters informed us about the adaptive potential of a population (the scale mutation rate of beneficial mutations, θ_{b}), the realized potential, as the number of strong beneficial under selection, and fitness diversity as the population genetic load. We applied this approach to a dataset of feral populations of honey bees (*Apis mellifera*) collected in California, and we estimated parameters consistent with the biology and the recent history of this species.

Keywords— Temporal data, Population genomics, Machine learning, Adaptation

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

One aim of population genomics is to understand how demography and natural selection shape the genetic diversity of populations. A classical approach is to assume that demography (migration, population subdivision, population size changes) leaves a genome-wide signal, whereas selection leaves a localizes signal close to where the causal mutation is located. Many methods follow this approach to infer demography or selection (reviewed by Beichman et al., 2018; Casillas and Barbadilla, 2017). Methods for demographic inference assume that most of the genome evolves without the influence of selection and that any deviation from the mutation-drift equilibrium observed in the data was caused by demographic events (Beichman et al., 2018). For selection, the majority of the methods search for locus-specific signals left by the beneficial mutation on nearby neutral mutations (Tajima, 1989; Fay and Wu, 2000; Kim and Nielsen, 2004) (low genetic diversity and high differentiation) to localize the beneficial mutation, assuming constant population size (Nielsen, 2005; Pool et al., 2010).

Conducting demographic and selection inference separately may have some shortcomings. First, there is the assumption that the signal left by demography is little affected by selection because selection is rare. However, linked selection can affect neutral and weakly selected sites that are far from the mutation targeted by selection (Sella et al., 2009; Neher, 2013) and selection can be pervasive (Sella et al., 2009; Lange and Pool, 2018). In addition, some methods for selection scans are not robust to misspecifications of demographic history. Consequently, an unspecified bottleneck or population increase, for example, can inflate the type I error rate of genome scans (Jensen et al., 2005, 2007; Schrider et al., 2016).These findings highlight the necessity of inferential methods that jointly accounts for the multiple evolutionary forces that act on populations (Lin et al., 2011; Li et al., 2012; Bank et al., 2014).

It is difficult to obtain the likelihood of models including demography and selection (but see Vitalis et al., 2014). Methods that rely on simulations provide alternatives to the use of likelihood functions (Csilléry et al., 2010; Schrider and Kern, 2018). One of the first works that proposed such strategy addressed the inference of local adaptation (Bazin et al., 2010). With coalescent simulations of an island model, Bazin et al. (2010) estimated demographic parameters and inferred the number of loci under selection. In their simulations, selection was modeled as differential locus-specific migration rates in which a selected locus had lower migration rates than neutrally-evolved loci. However, locus-specific migration rates or effective population size (as in Roux et al., 2016; Fraïsse et al., 2021) represent approximations of the selection process. Forward-in-time simulation allows more realistic models of selection. These were used to make inferences on N_e in the presence of selection by Sheehan and Song (2016) (for selective sweep and balancing selection) and Johri et al. (2020) (background selection). However, these works rely on simulations of few independent loci-not more than 50Kbp-which prevents the modelling of genome-wide effects of selection as the reduction of effective population size due to the variance of reproductive success of individuals (e.g. Santiago and Caballero, 1995). Nevertheless, this strategy brought new insights into the dynamics of selection. For instance, Laval et al. (2019) estimated the number of past selective sweeps that occurred in the human genome in the past 10,000, their intensity, and their age. Besides some limitations, these works exemplify the power of likelihood-free methods for the inference of the complex interaction between demography and selection.

Most population genetic studies use samples collected at one time point to infer the neutral processes (mutation, recombination, random genetic drift) and selection throughout the history of populations. Temporal data allows a better understanding of recent evolutionary processes (Feder et al., 2021; Dehasque et al., 2020) because they contain information about the allele frequency changes through time. By tracking the allele frequency changes over time, it is possible to estimate the relative role of selection and drift. Consequently, temporal data has the potential to give us a better understanding of the interaction between drift and selection (see for example,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Buffalo and Coop, 2019, 2020).

Here, we propose the use of ABC to jointly estimate demography and positive selection from temporal genomic data. In our framework, we use agent-based, forward-in-time simulations which allowed the modelling of the genome-wide and linked selection effects of beneficial mutations. Until recently, the use of such computationally demanding simulations in ABC inference was unrealistic since a great number of simulations are required to achieve accuracy in ABC (Frazier et al., 2018). However, with the introduction of Random Forests (ABC-RF), it was possible to reduce the computational burden as fewer simulations are required to achieve reliable estimates (Pudlo et al., 2016; Raynal et al., 2019). While many methods focus on the detection of targets of selection, our work addresses the inference of parameters that characterizes the genome-wide signal of demography and selection. Our genome-wide estimates showed to be reasonably accurate for a wide rate of adaptation and strength of selection. We were able to separate the estimates of N_{e}, which reflect the short-term drift from the population census size N. We also estimated the rate of influx of new beneficial mutations as measured by genome-wide scaled mutation rate of beneficial mutations. The separation between demography and drift, and the inference of genome-wide selection were only possible with the use of latent variables. Latent variables emerged as properties of each simulation and consequently, they better captured the emerging interaction between demography and selection than model parameters. We first evaluated the performance of an ABC-RF approach with forward-in-time simulations, and finally, we applied this framework to the analysis of a real time-series population genomics dataset of the feral population of honey bees (*Apis mellifera*) (Cridland et al., 2018). Our results were consistent with the biology of the species and with events that occurred recently in the history of the analyzed populations, taking into account the limitations of our approach.

Methods

Inference model

We assumed a closed population (no migration) of N diploid individuals that evolved under a Wright-Fisher model with selection. Selection only acted on *de novo* beneficial mutations and that were allowed to arise in the population since the first generation. Every beneficial mutation had a selection coefficient of s higher than zero and all were co-dominant. The values of the selection coefficients s were drawn from a gamma distribution with mean γ. Beneficial mutations entered the population with a rate of μ_{b} per generation that is independent of the mutation selective strength. Consequetly we defined the scaled mutation rate of the beneficial mutations per generation θ_{b} as the product the population size N, the mutation rate of beneficial mutation μ_{b} and the genome size G, $\theta_{\mathrm{b}} = 4N\mu_{\mathrm{b}}G$. This rate determines the amount of new beneficial mutations that arises in the population every generation. It can also be viewed as the waiting time for the appearance of a new beneficial mutation in the population. Populations with high θ_{b} receive new beneficial mutations every generation (Karasov et al., 2010), but population with low θ_{b} needs to wait more time for a new beneficial mutation to arise.

We needed to divide the model in two periods: 1) the burn-in period, that is necessary to remove, from the simulations, any footprint of the initial parameters set; the duration of this period was defined as the time necessary to contain all most recent common ancestors (MRCA) for all genomic regions in the simulation; and 2) the inference period, where we defined the longitudinal samples of individuals; the sampled genotypes were used to make inference of demography and selection. This two periods were defined by their time span and by the population census size with N_0 and N as the sample size of the burn-in and the inference period.

The first sample of individuals was taken at t_1, that was the first generation after the burn-in period ended, and the second was taken at t_2, after τ generations from t_1. Individuals were

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sampled following the sample plan II of Nei and Tajima (1981), where individuals were taken before their reproduction and permanently removed from the population. In this way their genotypes did not contribute to the next generation genotypes.

Each individuals genome of size G (in base pairs) consisted of a single linkage group with per base recombination rate per generation of c_0. In this genome, we modeled the act of selection by dividing it in "neutral" and "non-neutral" regions. Non-neutral regions held both neutral and beneficial mutations. This division can be interpreted as a genomic architecture in which there are genic regions that have a combination of neutral (synonymous intron mutations) and selected (non-synonymous mutation) sites, and intergenic regions (neutral mutations) but also other characteristics of the genome that would make the action of selection heterogeneous along the genome. We chose this simplification because it is general and it was a straightforward way to define independent priors for the relative number of non-neutral to neutral regions, and for the number of beneficial mutations in non-neutral regions. The probability of beneficial mutation to arise in the simulation, or their mutation rate per generation μ_{b}, was determined by the product of the proportion of non-neutral regions P_{R}, the proportion of beneficial mutation in a non-neutral region P_{B} and the mutation rate per generation μ. Figure 1 shows a schematic representation of the model template (and see Table S1 for a summary of the notation).

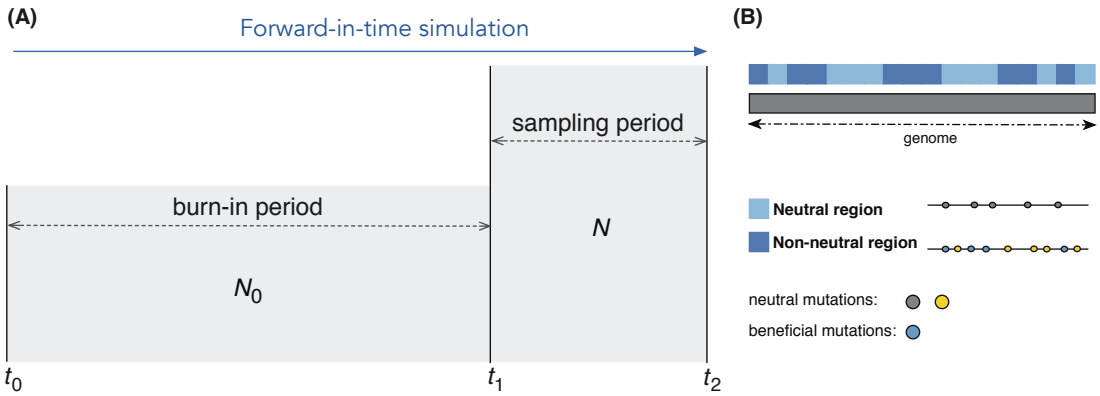

Figure 1: A schematic representation of the model used to simulate temporal population genomic data. (A) the population model that consisted of: 1) the burn-in period, where the number of generations was determined by the time necessary to contain the MRCA for all genomic regions. 2) the sampling period between the two time points, where the inference of demography and selection was made. (B) the genomic architecture model that consisted of: 1) a diploid genome of one linkage group that was divided in neutral and non-neutral regions that were composed of neutral and a combination of neutral and beneficial mutations.

Calculation of summary statistics and latent variables

The above model was used to simulate the dynamic of drift and selection in a closed population. In the two sample periods, individuals were sampled and used for the calculation of the summary statistics for the ABC-RF framework. For each simulation, we calculated summary statistics that: 1) compared the two samples (*e.g.* genetic differentiation F_{ST}), and 2) quantified the diversity within-sample (*e.g.* expected heterozygosity H_{E}). For the later, calculations were obtained for each sample and the pooled sample (when the genetic data of populations were combined in one population). Some summary statistics were calculated genome-wide, for example, global F_{ST}, global H_{E} and the total number of polymorphic sites S; others were calculated

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNP-by-SNP as the H_{E}; or they were calculated in windows as S, the nucleotide diversity π, and Tajima's D. For every simulation, we measured the mean, variance, kurtosis, skewness, and 5% and 95% quantiles among all locus-specific or window summary statistics. These statistics informed about the heterogeneity of genome-wide distribution of locus-specific or window summary statistics. We set three window sizes for the window summary statistics: 500, 5,000, and 10,000 bp. Those windows overlapped because each window was composed around every single SNP which put the targeted variation in the middle of the window with the other SNPs in half of the window size on each side of the targeted SNP. The site-frequency spectrum was obtained as a global summary statistics with three different numbers of discrete classes (bin sizes): 10, 15, and 20 bins (the complete list of summary statistics can be found in Supplementary Methods, section S2.1 List of summary statistics).

For every simulation, we combined a vector of summary statistics with the vector of X model parameters and the vector of five latent variables. Latent variables represents values that emerged from the simulation or values that emerged by combining a latent variable and a model parameter. In our inferential framework, for example, the effective population size N_{e} is a latent variable and it was calculated within each simulation. The ratio between the effective population size N_{e} and the population census size N, N_{e}/N, on the other hand, was derived by combining a latent variable and a model parameter for each simulation. The other three latent variables were: the number of beneficial mutations under strong selection P, the average selection coefficient of strongly selected mutations $\bar{s}$, and the average substitution load L.

The effective population size N_{e} measures the increase of inbreeding at each generation. In this definition, N_{e} is the size of an ideal population with the same amount of drift as the population under consideration. Defined in these terms, N_{e} is the inbreeding effective size (Santiago and Caballero, 1995; Walsh and Lynch, 2018). It was calculated in every generation i of the sampling period as:

$$N_{\mathrm{e},i} = \frac{4N}{\sigma^2_{k_i} + 2}$$

$\sigma^2_{k_i}$ being the variance among parents of the number of gametes produced that contributed to offspring in generation i. The N_{e} for the whole inference period was obtained by calculating the harmonic mean of $N_{\mathrm{e},i}$. The population size of N was kept constant for the whole period as shown above, and it represents a parameter of the simulation. From the N_{e} we obtained the ratio N_{e}/N (it measures how the census size reflects the actual effective population size: we expect to have a reduction on N_{e} compared to N when beneficial mutations are more pervasive).

We also recorded the selection coefficient of all beneficial mutations that arose in every generation i from t_1 to t_2 in each simulation. After, we calculated the fraction of beneficial mutations that were strongly selected (where $s > 1/N_{\mathrm{e}}$ over all mutations that were segregating in the period). This fraction represented all beneficial mutations that arose in between t_1 and t_2, regardless if they were lost or fixed at any generation of the period, or if their frequency oscillated but never reached fixation. We decided for it because any beneficial mutation that arise can impact the allele frequency trajectories of other mutations (neutral or beneficial). For these mutations, we also calculated the average across all selection coefficients. We also calculated, in every generation of this period, the substitution load L_i as the difference between the total fitness of the individual with the highest fitness $W_{\mathrm{max}i}$ and mean total fitness of the population $\bar{W}_i$ (it measures the overall diversity of beneficial mutations present in the inference period),

$$L_i = \begin{cases} 0, & \text{if } W_{\mathrm{max}i} = 0 \\ \frac{W_{\mathrm{max}i} - \bar{W}_i}{W_{\mathrm{max}i}}, & \text{otherwise} \end{cases}$$

The average substitution load was obtained by averaging all values of L_i.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Implementation

The model was simulated with the software SLiM v3.1 (Haller and Messer, 2017; Haller et al., 2019). For the calculation of the inbreeding effective size, we needed to activate an optional SLiM 3.1 behavior to track the pedigrees of each individual in the population. It allowed us to obtain the number of each parent gamete and the population variance of the number of gametes. For the calculation of the generation substitution load, we used a SLiM built-in function that allowed us to obtain the vector of fitness of all individuals in the population. The cached fitness was the sum of all fitness determined by each beneficial mutation.

Each simulation was produced by using different combinations of the model's parameters: 1) the mutation rate per bp per generation μ, 2) the per-base recombination rate per generation c_0, 3) the mean γ of a gamma distribution with the shape parameters equal to the mean, from which the selection coefficients s of each beneficial mutation in the simulation were sampled, 4) the number of non-neutral genomic regions P_{R}, 5) the parameter that determines the probability of beneficial mutation in non-neutral regions P_{B}, 6) the population census size of the burn-in period N_0, and, finally, 7) the population size of the inferential period N.

We set SLiM to output genotypic data of samples of individuals as single nucleotide polymorphisms (SNPs), at t_1 and t_2, in the VCF file format. Using bcftools (Li, 2011), custom R function (R Core Team, 2020) and EggLib (De Mita and Siol, 2012), SLiM outputs were processed and summary statistics calculated. We implemented a pipeline in an R script that automates the sampling of the prior values, runs each simulation, manipulates the VCF files, calculates the summary statistics, and organizes the final reference table. This script was also produced to facilitate the model test with few simulations and the job submission in a HPC node(s). The main R and additional scripts are available on Zenodo (Pavinato et al., 2021). In this pipeline, for every simulation, a row of the reference table was produced by combining the model parameters used to launch a SLiM simulation, latent variables, and the summary statistics.

ABC-RF

In this work, we take advantage of the use of Random Forests (RF) in ABC procedure, where the parameter estimation is a machine learning problem (Pudlo et al., 2016; Raynal et al., 2019).

A reference table was generated with the model described above. In each simulation, each individual had a genome of size 100 Mb that was divided into 2,000 fragments of 50,000 bps. A number of these fragments were randomly set as either neutral or non-neutral, based on the parameter P_{R}. In the inference period, 100 individual genotypes were sampled at t_1 and t_2 after $\tau = 10$ generations. For all model parameters, values of each simulation were sampled from a log-uniform distribution with range: 1 to 2,000 for N_0 and N, 10^{-10} to 10^{-6} for μ, 5×10^{-10} to 5×10^{-7} for r, 10^{-5} to 1 for P_{B}, and 10^{-3} to 1 for γ. And uniform distribution with range 0 to 1 for P_{R} (Figure S1 shows the prior distribution for all model parameters and latent values).

The raw reference table produced by the pipeline was processed to remove missing data. Missing data were present in several summary statistics of simulations with low genetic diversity, that can be produced, for example, by low mutation rate, by small population size, by selection, or by the combination of these parameters. Missing data were also present in the entire row of a simulation if the combination of population size, mutation, and especially recombination rate produced simulations that were memory intense, which caused the simulation to crash. A final reference table containing 55,634 simulations with 405 summary statistics was used to train the ABC-RFs. Independent RFs were obtained for each parameter and latent variable using R package abcrf (Pudlo et al., 2016; Raynal et al., 2019). Each RF was obtained by growing 1,000 trees. The RF were grown with the default parameters. Average genetic load, L and P were logit transformed prior the training. For these latent variables and for $\bar{s}$, simulations with $L = 0$, $P = 0$ or $\bar{s} = 0$ were also excluded from the training set, which reduced it to 36,026

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

simulations for L, and with 29,264 simulations for P and $\bar{s}$. For the other parameters and latent variables, we performed log transformation prior training and use the reference table containing all simulations.

The performance of each trained Random Forest was evaluated with *out-of-bag* (OOB) estimates (Breiman, 2001). These estimates were produced by the trained model for the data used for training. Regression trees that compose the actual RF are grown using part of the data, selected randomly from the initial set of simulations. As a consequence, for each simulation, there is a subset of trees that were grown without the data from that simulation. The estimate from that subset of trees is called OOB estimate and with it, cross-validation of the trained model is done without splitting the reference table in the training and testing sets. We calculated the mean squared error (MSE) and the correlation coefficient (R^2) between the true and the OOB estimated values obtained with the function regAbcrf implemented in the R package abcrf.

Alternative estimates of N_e from temporal data

We compared the the ABC-RF N_e estimates with estimates obtained with the global F_ST between temporal genomic samples (Frachon et al., 2017). This estimator is defined as:

$$\hat{N}_\mathrm{e} = \frac{\tau(1 - \hat{F}_\mathrm{ST})}{4\hat{F}_\mathrm{ST}}$$

where τ accounts for the time-interval, in generations, between the first and the last samples used to estimate the F_ST, and $\hat{F}_\mathrm{ST}$ is the the Weir and Cockerham's F_ST estimator (Weir and Cockerham, 1984). The N_e from the F_ST was calculated for all simulations used to train the random forest. We calculated the mean squared error (MSE) and the squared correlation coefficient of linear regression (R^2) between the observed (true) and the F_ST-based N_e estimated values of all simulations. We also evaluated the performance of each estimator by calculating the MSE for simulations within a specific range of values of θ_b (local MSE estimates). By comparing the changes in MSE values of each estimator as a function of θ_b we could better understand how the amount of selection affected each estimator.

Analysis of temporal genomic data of feral populations of *Apis mellifera*

To give an example of an application of our framework, we analyzed a subset of whole-genome sequencing data of feral populations of honey bees from California (Cridland et al., 2018). Eight out of fourteen sites in this work were composed of samples from museum and contemporary collections of freely foraging honey bees: 1) Avalon site in Catalina Island, Los Angeles county, 2) Arcata and Blue Lake sites in Humboldt county, 3) Placerita Canion Nature Area in Los Angeles county, 4) Sky Valley and Idyllwild in Riverside County, 5) La Grange, Stanislaus county, 6) Stebbins Could Canyon Reserve, Solano county and 7) UC Davis Campus, Yolo county. This dataset contains pairs spanning 104 years (as in the Avalon site, Catalina Island, Los Angeles county) and pairs spanning only 15 years (as in the Placerita Canyon Nature Area, Southern California, and Idyllwild, in Riverside county). For the temporal samples from Riverside County, we only used the two samples collected in May 1999 in Idyllwild as the first sample, and combined all samples collected in September 2014 (in Idyllwild and Sky Valley) as the second sample.

Individual VCF files of each site were combined with bcftools (Li, 2011) and a custom R script was used to convert the dataset to an input format required to run an EggLib custom implementation (in Pavinato et al., 2021). We tagged samples from the same time point with the same label. For the simulations, we set the genome size of each individual to 250 Mb (similar to the most recent estimates of *A. mellifera* genome size (Elsik et al., 2014). For the analysis

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of the *Apis mellifera* temporal genomic data we measured the amount of missing data, but we found a negligible amount ($< 1\%$) in most of populations (except populations from Avalon and Placerita that had 10% of the total missing genotypes).

The genome was divided into 5,000 fragments of 50,000 bps. These fragments were randomly set as neutral or non-neutral according to the parameter P_{R}. Dominance coefficients were set to 0.5 for all beneficial mutations throughout the simulation. In the sampling period, for each site, the same number of individuals found in each dataset were sampled in the first (t_1) and second time point (t_2). We used a Normal distribution for μ with a mean 3.4×10^{-9} with a standard deviation of 0.5, to have a prior distribution center around the estimated mutation rate for Hymenoptera (Liu et al., 2017). The per base recombination rate was set as Uniform, ranging from 10^{-8} to 10^{-4}. The genome was represented with a single linkage group. For the population sizes N_0 and N were taken from a Uniform prior distribution ranging from 1 to 10,000 individuals. Other prior probability distribution of the parameters were set with the same prior as described above. We used the same summary statistics as described above, except this time, we only calculated one window size of 10Kbp for summary statistics calculated in windows and one bin size of 10 bins for the site-frequency spectrum. The raw reference table containing the vector of parameters, latent variables, and summary statistics produced by the pipeline was processed to remove missing data, and a final reference table containing 162 summary statistics for each population pair, was used to train the ABC-RFs. (see Table 1 for the number of simulations of each reference table). The RFs were grown as described above.

Table 1: Populations and number of simulations in the reference table.

Location	Date	N	Simulations
Avalon, Catalina Island, Los Angeles county	1910/2014	2,5	13,953
Blue Lake and Arcata, **Humboldt** county	1966/2015	6,6	14,216
Placerita Canyon Nature Area, Los Angeles county	1999/2014	5,6	14,125
Idyllwild and Sky Valey, **Riverside** county	1999/2014	2,8	13,930
La Grange, **Stanislaus** county	1976/2014	2,6	13,956
Stebbins Cold Canyon Reserve, Los Angeles county	1996/2014	5,5	14,121
UC **Davis** Campus, Yolo county	1968/2015	2,6	13,970

Names highlighted in bold letters corresponds to the population code we used in this work.

Results

ABC-RF framework for joint inference of adaptive and demographic history

The ABC-RF framework jointly estimated parameters informative about adaptive and demographic history in temporal population genomics settings. Independent random forests estimated the population scaled beneficial mutation rate θ_{b}, the population census size N, and the effective population size N_{e} (Figure 2). Trained RFs performed well in predicting N and N_{e} with small MSE and higher R^2 (Figure 2 b and c). But, the trained RF for θ_{b} had a lower performance than the trained RFs for demographic parameters, with high MSE and low R^2 (Figure 2a and b). Still, the estimates were robust for intermediate to higher values of θ_{b}. For the results of other model parameters and latent variables informative about demography and selection, see S3 Supplementary Results.

The automated selection of informative summary statistics is an important feature of ABC-RF. For each tree of a random forest, summary statistics were selected given its ability to split the data. How many times a summary statistics was selected in each RF informs us of their

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

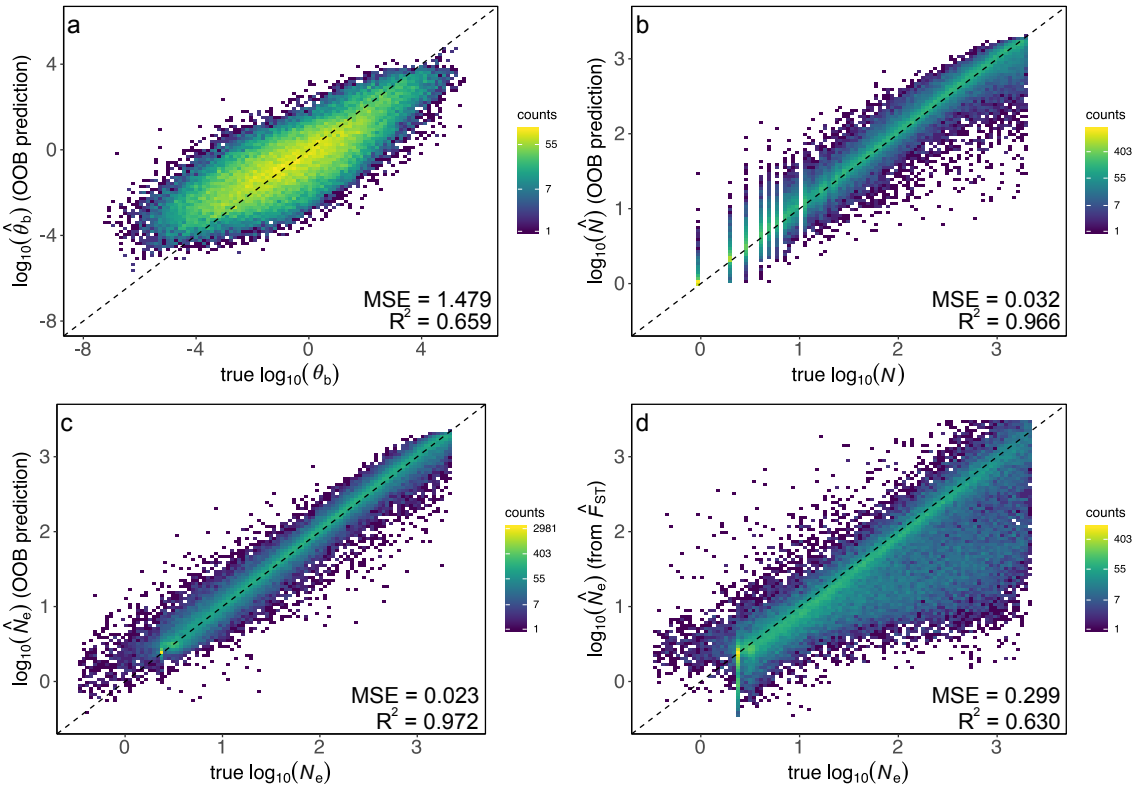

Figure 2: Out-of-bag estimates of ABC-RF trained for the joint inference of demography and selection, and $\hat{N}_e$ estimates from the temporal F_{ST} to compare with the ABC-RF -based $\hat{N}_e$ estimates. (a) population census size N; (d) effective population size N_e; (c) population scaled mutation rate of beneficial mutations θ_b; and (d) N_e from temporal F_{ST}

importance for the inference of a given parameter. For the prediction of θ_b values, the RF picked more frequently the 5% quantile of Tajima's calculated in the second sample, with the kurtosis and skewness of F_{ST} and D_a calculated globally (Figure S4 e). The population size was trained with a combination of within and between sample summary statistics: F_{ST} and D_a, with their respective derived statistics were frequently selected (Figure S5 c). For N_e, summary statistics that inform about the cumulative divergence between samples as F_{ST} and D_a, were frequently selected (Figure S5 d).

Comparison with F_{ST} method to estimate N_e

We compared our ABC-RF N_e estimates with estimates obtained with the temporal F_{ST} (Frachon et al., 2017). The F_{ST}-N_e was more affected by the amount of selection in larger populations. Consequently, the F_{ST}-N_e showed higher overall MSE and lower R^2 compared to the ABC-RF -based estimates (Figure 2c and d). When the beneficial mutations were less frequent (either because of the mutation rate of the beneficial mutation, and because of larger population size), the ABC-RF and the temporal F_{ST} performed well and similarly regardless the strength of selection, with the ABC-RF based estimator with less error than the temporal F_{ST}-based estimator. However, when the frequency of selection started to increase, the N_e estimator based on the temporal F_{ST} had dramatically higher error (Figure 3).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

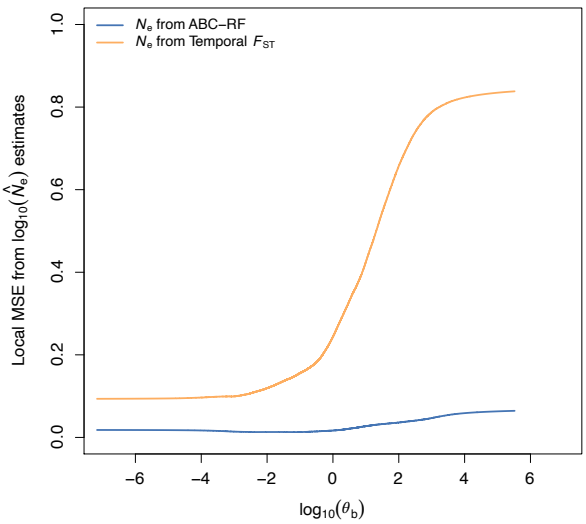

Figure 3: Local MSE of N_e estimates as a function of θ_b. The lines corresponds to the MSE on N_e estimates from ABC-RF and from temporal F_{ST}.

Analysis of temporal genomic data of feral populations of *Apis mellifera*

For the analysis of feral *A. mellifera* populations, we first grew independent RF for each parameter in each population. All populations had the same performance of the ABC-RF estimator for N_e, as they showed similar values of MSE and R^2 (Figure S8). For N, trained RF for Humboldt, Stebbins and Placerita performed similarly well, and they had the lowest MSE and higher R^2 (Figure S7). For θ_b, Riverside had trained RF with the worst performance (Figure S6). Overall, both MSE and R^2 obtained with OOB estimates from simulated data for *Apis mellifera* dataset were comparable to these parameters obtained with OOB estimates for the simulated data for the evaluation of the method.

Trained RF for N and N_e were able to predict these parameters in all populations, as the inference of the mean posterior value and the posterior distribution differentiated from the mean prior value and distribution (Figure 4 b and c). For N, posterior distribution were wider than for N_e. Trained RF for θ_b, for all populations had a similar posterior mean, except for the Avalon population that had a peak at a lower value (Figure 4 a). But, the posterior distributions were wider and they followed the prior distribution, making it difficult to accurately predict the posterior mean and variance in all populations. It is possible to see together with the posterior mean estimates that the ABC-RF estimates for θ_b were concentrated in lower values (Table S2) in all populations. N_e were also lower, and N_e and N were similar. For the results of OOB estimates of other model parameters and latent variables, and for posterior estimates for these parameters see S3 Supplementary Results.

Discussion

Separating demography from drift, and the inference of θ_b

With temporal population genomics data, we can see the evolution in "action" as opposed to single time-point population genomics data (Feder et al., 2021). Consequently, temporal data have more information about the ongoing process, which make them better for the understanding of the complex interaction between demography and selection (Buffalo and Coop, 2019; Dehasque et al., 2020; Williams and Pennings, 2020). When samples from more than two time points are available, the footprints of drift and selection in allele frequencies changes can

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

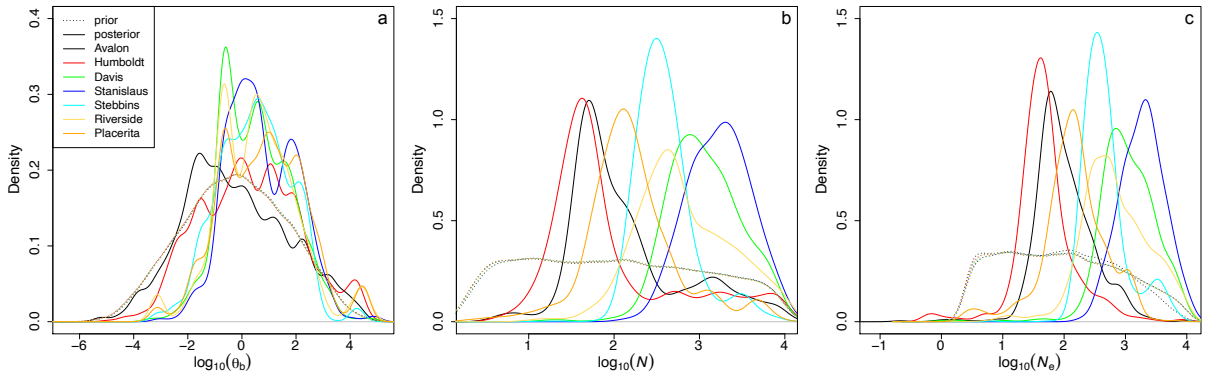

Figure 4: Joint inference of demography and selection for feral *A. mellifera* populations. (a) the population census size N, (b) the effective population size N_e (c) the scale mutation rate of selected mutations θ_b. Dashed and filled lines corresponds to the prior and posterior distributions, respectively

be separated (e.g., Buffalo and Coop, 2020; Feder et al., 2014). Our results showed that two samples collected in different time points were sufficient for the inference of the genome-wide footprint of adaptive evolution and to separate the demography (population census size N) from drift (effective population size N_e).

It is important to stress that N_e, as a latent variable, captures the feedback dynamics between drift and linked selection. Selection, either positive or negative, causes a deviation of N_e from N. The impact of selection on the genome can extend far from the target of selection since individuals that carry beneficial mutations have more chance to reproduce, their beneficial mutations are more likely to be in the next-generation offspring (Walsh and Lynch, 2018). In this complex dynamic, with many loci under selection which creates a dynamic that cannot be easily described, latent variables actually summarized the by product of drift and selection interactions. With our approach, $\hat{N}_e$ quantifies the drift due both to demographic and selection dynamics.

This genome-wide reduction in N_e is not captured when loci are assumed to evolve independently (as in Sheehan and Song, 2016, for example). In contrast, the complexity of linked selection and the genome-wide effect of selection are taken into account by using individual based simulations with whole genome in an ABC approach.

Estimates of genetic load or other genome-wide parameters about selection are obtained when annotated genomic data is available, (Henn et al., 2015) or by conducting experiments on crossing populations (for the genetic load Plough (2016)). However, we were able to obtain estimates of selection parameters only with polymorphism data. A similar approach was used by Laval et al. (2019) to estimate the number of selective sweeps affecting the human genome in the past 10,000 years. But they did not consider the feedback dynamics of selection and demography. Differently, Buffalo and Coop (2020) measured the genome-wide signature of selection by estimating the covariance of allele frequencies at consecutive time points. This allowed the quantification of the genome-wide contribution of selection to the observed allele frequency changes, even when selection involved many loci of small effect. In this work, we estimated the population scale mutation rate of beneficial mutations θ_b, which informs about the diversity of beneficial mutations that existed in the population between the two time points, and the potential speed of adaptation at the genome level (Hermisson and Pennings, 2017). These estimates reflect the potential number of beneficial mutations present between the two time points regardless their impact as determined by their selection coefficients.

The variable importance plot of each parameter shows us the global importance of each

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

summary statistics in the trained Random Forests. For N_e, N, and θ_b summary statistics calculated from the distribution of locus-specific summary statistics -skewness, kurtosis, mean, variance, 5% and 95% quantiles were more frequently used. Summary statistics derived from the distribution of locus-specific calculated from all segregating loci in the genome inform about the heterogeneity that selection and drift produce genome-wide. For example, a *de novo* a beneficial mutation entered the simulation and was selected, it left a signal of lower diversity around the region it was located. The genome, after selection, contained spots where diversity was high and where it was low, and this heterogeneity was captured by the distribution of locus-specific H_E, more specifically the lower tail of the distribution where the values of diversity were lower. The use of the covariance matrix of allele frequencies through Buffalo and Coop (2020) can be considered as a summary statistic for the the genome-wide heterogeneity that selection and drift left on the genome. It would be interesting to include this matrix as summary statistics for further development of the method.

Comparison with F_{ST} method to estimate N_e

We compared the N_e obtained with ABC-RF framework to the N_e obtained with F_{ST} estimator (Skoglund et al., 2014; Frachon et al., 2017). Overall, the F_{ST} -based N_e estimator performed poorly compared to the ABC-RF -based estimator. The lower performance were caused by N_e values that were underestimated when beneficial mutations were more frequent (higher θ_b). Consequently, the N_e estimates from the temporal F_{ST} were strongly affected by selection. When selection was infrequent or rare, both estimators performed similarly well, but with the ABC-RF estimator with lower MSE. Positive selection can increase the variance of allele frequency between samples taken in different time points. When selection is infrequent or rare, drift determines most of allele frequency changes between samples; but when selection is pervasive, selection dominates, which cause dramatic and rapid changes in allele frequency, increasing the variance between samples. N_e estimator based on the F_{ST} depends on the differences in allele frequencies between samples, consequently it is naturally biased by strong and frequent selection. We can assume that the N_e estimator from ABC-RF was insensitive to the amount of selection since we trained the ABC-RF with N_e values from the simulation. In our simulations, N_e was a latent variable that captured the deviation that selection imposed on the number of individuals able to reproduce (selected for); it was not a biased by unaccounted factors.

The amount of selection for $\theta_b \sim O(1)$ could be unrealistic in some organisms, but plausible in virus (Feder et al., 2014) and many arthropod species, with large N_e, which have larger population sizes (except in eusocial insects that have vertebrate-like population sizes Romiguier et al. (2014)). In larger populations sizes selection acts also on weaker and milder beneficial mutations. In those organisms, it might be unreasonable to assume mutation-drift equilibrium given the pervasive role of selection. Consequently, any attempt to estimate demography parameters as N_e without properly accounting for the pervasive role of selection could be biased.

Analysis of temporal genomic data of feral populations of *Apis mellifera*

Overall, the performance of the ABC-RF for selection and demography inference was similar across populations despite the differences in sample size and age. For θ_b, Avalon and Humboldt populations had posterior probability distributions very similar to the prior, indicating that the analysis provides no additional information on this parameter. These two population also present low effective population size estimates, which can reduce the signal of selection. For the rest of the populations, the posterior probability distribution of θ_b is tilted toward the higher values but without a clear peak differentiating the distribution from the prior. Still, lower θ_b values could be excluded. This favours an interpretation in which selection was acting during the study period but without providing a precise estimate of the parameter. The information about the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

presence of selection in these analyses comes mainly from the heterogeneity of the polymorphism along the genome, thus, for a thorough interpretation of the results, it is important to discuss other processes that have not been modelled but that could affect this signal. The studied bee populations in California show a mixture of Eastern and Western European ancestry, with some populations presenting African ancestry in the most modern samples Cridland et al. (2018). Different levels of African admixture along the genome could create some heterogeneity and affect the inference. However, Placerita and Riverside, the populations with higher African ancestry at present, present similar estimates of θ_b that populations with little or no African admixture. Also, Humboldt population changes from having predominately Western European ancestry to having predominately Eastern European ancestry, which means that there was substantial gene flow into the population. These results suggest that admixture does not dramatically affect the inference of selection but also highlights the importance of incorporating admixture in future development of the approach.

Our ABC-RF approach estimated N_e with the same order of magnitude of other N_e estimates obtained for hymenopterans (Zayed, 2004). Lower values of N_e might reflect the presence of admixture, either African admixture or admixture that occurred with domesticated lineages facilitated by changes in beekeeping practices. Northern populations, especially from Humboldt County, shared similarities with bees from reared colonies (with higher Eastern European ancestry). Southern populations, as shown by Cridland et al. (2018) showed a higher level of admixture with African lineages. Populations from the southernmost cites (from Riverside County, Placerita, and Avalon, Los Angeles county) showed higher genetic diversity, but they did not show the highest values of N_e. On the other hand, the population of Stanislaus County had the highest value of N_e, possibly because it had lower levels of admixture with domesticated lineages compared to the population from Riverside, Placerita, Avalon, and Los Angeles county.

We observed that N_e and N had similar estimates. We were aware that our simulation model did not account for key characteristics of eusocial insect reproductive biology: the monopolization of reproduction by the queen and the division of labor. In Honey bees, a queen mates with more than one male (a process called polyandry) (Estoup et al., 1994), which leads to a biased breeding sex ratio. Assuming that only queens can reproduce in the colony, polyandry increases the variance in the number of parents that contribute to the offspring gene pool, which leads to a decrease in the N_e compared respect to N (Nomura and Takahashi, 2012). In our simulations, we only simulated monogamous mating, therefore, the difference between estimates of N_e and N only reflects the action of selection. Therefore N must be interpreted with caution as it is probably reflecting more the total number of female breeders per generation rather than the size of the population. Individual-based forward simulators as SLiM allows setting different mating schemes. It is possible to simulate the haplodiploidy, the cast system, diocy and sex ratio found in honey bees. These modifications in the simulation could potentially allow us to estimate N and other parameters that could reflect better the biology of the species but it was not the focus of this work.

One possible explanation for the similarities between N_e and N estimates, thus, relies on cast specialization and concentration of reproduction to one of few females in the colony. These came to a cost of reduced N_e, which reduces the efficacy of selection (either positive or negative). Bees are the few insect groups that show very small N_e potentially linked with the evolution of eusociality (Romiguier et al., 2014). Knowing that lower N_e reduces the effectiveness of selection, it is plausible to think that lower N_e is restricting the effects of mutation affecting fitness to stronger beneficial mutations. Since these mutations are less frequent than weak or mild mutations, their effects on N_e were small, which explains why N_e and N had values in the same range. Low N_e and low θ_b pointed to a biological system limited where adaptation is limited by the influx of adaptive mutations (Rousselle et al., 2020).

Our ABC-RF framework also estimated the per-site mutation rate per generation μ (Supplementary Results, S18). For all populations, the mean posterior μ exceeds the mean μ of

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the species. The higher estimated values we obtained might reflect the recent admixture events between these populations. We did not model gene flow and admixture. Modeling admixture could help us correctly separate the effects of selection and drift since the introgression of African genes might have biased some estimates of selection parameters.

Perspectives and Limitations

Our model is very simplistic, as it only considered the impact of beneficial mutations, neglecting the effect of background selection and standing variation. Background selection can mimic directional selection because they cause a similar pattern of diversity reduction around the target of selection. However, it was recently shown that background selection only mimics the classical sweep in simplistic models, where the deleterious mutation is localized in a specific region of the genome (Schrider, 2020). For more realistic scenarios, where the concentration of deleterious mutations varies across the genome, background selection does not behave as a classical hard sweep. In an attempt to jointly accommodate the effect of demography and selection on the inference of N_e, Johri et al. (2020) modeled the effect of background selection and developed an ABC-based approach that jointly estimated the distribution of fitness effects and N_e. In their simulations, they modeled deleterious mutations and the classical hard sweep with the inclusion of beneficial mutations. They showed an unbiased estimate of N_e regardless of the presence of positive and negative selection. Future developments should include a more realistic genomic architecture where both negative and positive mutations can co-occur and explore different concentrations of deleterious mutations. In addition to that, further developments should explore not only scenarios of *de novo* mutations, but selection acting on standing variation. This can be easily achieved with our pipeline and allows for a more general treatment of the selection of soft sweeps. The model can also be expanded to more complex demographic scenarios, including changes in population size and genetic exchange with external sources (migration). Including such admixtures will be key in the future development of this approach since it is also a source of heterogeneity in the genome and, thus, might influence the performance of the method.

Conclusion

We show that an ABC-RF -based approach is able to jointly infer adaptive and demographic history from temporal population genomics data. This approach allows the quantification of the genome-wide footprint of selection expressed in the scaled mutation rate of beneficial mutations. The ABC-RF N_e is robust to varying degrees of strength of selection and frequency of beneficial mutations. Our ABC-RF -based approach can be applied to temporal population genomics datasets to gain insight about the adaptive and demographic history of natural populations.

Acknowledgements

This project has received funding from the LabEx AGRO (convention ANR-10-LABX-0001-01), CEMEB (convention ANR-10-LABX-0004) and NUMEV (convention ANR-10-LABX-20) through the AAP Inter-LabEx (ABCSelection). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 791695 (TimeAdapt). S. De Mita was funded by INRAE (*Projet Innovant EFPA*). We are thankful for the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12), for the High-Performance Computing Center University of Montpellier (MESO@LR-Platform) and for the Ohio Super Computer Center (OSC Center (1987)) for providing computing resources. The authors would like to thanks Julie Cridland for sharing the processed data, and Andrew P. Michel for providing suggestions to improve an early version of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., and Jensen, J. D. (2014). Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, 30(12):540–546.

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

Beichman, A. C., Huerta-Sanchez, E., and Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 49(1):433–456.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Buffalo, V. and Coop, G. (2019). The linked selection signature of rapid adaptation in temporal genomic data. *Genetics*, 213(3):1007–1045.

Buffalo, V. and Coop, G. (2020). Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences of the United States of America*, 2:201919039.

Casillas, S. and Barbadilla, A. (2017). Molecular population genetics. *Genetics*, 205(3):1003–1035.

Center, O. S. (1987). Ohio supercomputer center.

Cridland, J. M., Ramirez, S. R., Dean, C. A., Sciligo, A., and Tsutsui, N. D. (2018). Genome sequencing of museum specimens reveals rapid changes in the genetic composition of honey bees in California. *Genome Biology and Evolution*, 10(2):458–472.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and Francois, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25(7):410–418.

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

Dehasque, M., Ávila-Arcos, M. C., Díez-del Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E. D., Malaspinas, A.-S., Marques-Bonet, T., Martin, M. D., Murray, G. G. R., Papadopulos, A. S. T., Therkildsen, N. O., Wegmann, D., Dalén, L., and Foote, A. D. (2020). Inference of natural selection from ancient DNA. *Evolution Letters*, 29:753–15.

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J. D., Foster, L. J., Graur, D., Guigo, R., HGSC production teams, Hoff, K. J., Holder, M. E., Hudson, M. E., Hunt, G. J., Jiang, H., Joshi, V., Khetani, R. S., Kosarev, P., Kovar, C. L., Ma, J., Maleszka, R., Moritz, R. F. A., Munoz-Torres, M. C., Murphy, T. D., Muzny, D. M., Newsham, I. F., Reese, J. T., Robertson, H. M., Robinson, G. E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J. M., Vaerenbergh, M. V., Waterhouse, R. M., Weaver, D. B., Whitfield, C. W., Wu, Y., Zdobnov, E. M., Zhang, L., Zhu, D., Gibbs, R. A., and Honey Bee Genome Sequencing Consortium (2014). Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics*, 15:86.

Estoup, A., the, M. S. P. o., and 1994 (1994). Precise assessment of the number of patrilines and of genetic relatedness in honeybee colonies. *Proceedings of the Royal Society B*, 258:1–7.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fay, J. C. and Wu, C.-I. (2000). Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413.

Feder, A. F., Kryazhimskiy, S., and Plotkin, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522.

Feder, A. F., Pennings, P. S., and Petrov, D. A. (2021). The clarifying role of time series data in the population genetics of HIV. *PLoS Genetics*, 17(1):e1009050.

Frachon, L., Libourel, C., Villoutreix, R., Carrère, S., Glorieux, C., Huard-Chauveau, C., Navascués, M., Gay, L., Vitalis, R., Baron, E., Amsellem, L., Bouchez, O., Vidal, M., LE CORRE, V., Roby, D., Bergelson, J., and Roux, F. (2017). Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nature Ecology & Evolution*, 1(10):1551–1561.

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., LOIRE, E., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., and Roux, C. (2021). DILS: Demographic inferences with linked selection by using ABC. *Molecular Ecology Resources*.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19(2):552–566.

Haller, B. C. and Messer, P. W. (2017). SLiM 2: flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1):230–240.

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., and Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16(6):333–343.

Hermisson, J. and Pennings, P. S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution*, 8(6):700–716.

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., and Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics*, 170(3):1401–1410.

Jensen, J. D., Thornton, K. R., Bustamante, C. D., and Aquadro, C. F. (2007). On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, 176(4):2371–2379.

Johri, P., Charlesworth, B., and Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192.

Karasov, T., Messer, P. W., and Petrov, D. A. (2010). Evidence that adaptation in drosophila is not limited by mutation at single sites. *PLoS Genetics*, 6(6):e1000924.

Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524.

Lange, J. D. and Pool, J. E. (2018). Impacts of recurrent hitchhiking on divergence and demographic inference in Drosophila. *Genome Biology and Evolution*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Laval, G., Patin, E., Boutillier, P., and Quintana-Murci, L. (2019). A genome-wide Approximate Bayesian Computation approach suggests only limited numbers of soft sweeps in humans over the last 100,000 years. *bioRxiv*, 5:2019.12.22.886234.

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics*, 27(8):1157–1158.

Li, J., Li, H., Jakobsson, M., Li, S., Sjödin, P., and Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, 21(1):28–44.

Lin, K., Li, H., Schlötterer, C., and Futschik, A. (2011). Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, 187(1):229–244.

Liu, H., Jia, Y., Sun, X., Tian, D., Hurst, L. D., and Yang, S. (2017). Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. *Molecular Biology and Evolution*, 34(1):119–130.

Neher, R. A. (2013). Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):195–215.

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

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1):197–218.

Nomura, T. and Takahashi, J. (2012). Effective population size in eusocial Hymenoptera with worker-produced males. *Heredity*, 109(5):261–268.

Pavinato, V., De Mita, S., and Navascués, M. (2021). Tracking-selection. doi:10.5281/zenodo.4599736.

Plough, L. V. (2016). Genetic load in marine animals: a review. *Current Zoology*, 62(6):567–579.

Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Research*, 20(3):291–300.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J. M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

Romiguier, J., Lourenco, J., Gayral, P., Faivre, N., Weinert, L. A., Ravel, S., Ballenghien, M., Cahais, V., Bernard, A., Loire, E., Keller, L., and Galtier, N. (2014). Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *Journal of Evolutionary Biology*, 27(3):593–603.

Rousselle, M., Simion, P., Tilak, M.-K., Figuet, E., Nabholz, B., and Galtier, N. (2020). Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLoS Genetics*, 16(4):e1008668.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.12.435133; this version posted October 18, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biol*, 14(12):e2000234–22.

Santiago, E. and Caballero, A. (1995). Effective size of populations under selection. *Genetics*, 139(2):1013–1030.

Schrider, D. R. (2020). Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics*, 216(2):499–519.

Schrider, D. R. and Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4):301–312.

Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204(3):1207–1223.

Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? *PLoS Genetics*, 5(6):e1000495.

Sheehan, S. and Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3):e1004845.

Skoglund, P., Sjödin, P., Skoglund, T., Lascoux, M., and Jakobsson, M. (2014). Investigating population history using temporal genetic differentiation. *Molecular Biology and Evolution*, 31(9):2516–2527.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

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

Walsh, B. and Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-Statistics for the analysis of population structure. *Evolution*, 38(6):1358.

Williams, K.-A. and Pennings, P. (2020). Drug resistance evolution in HIV in the Late 1990s: hard sweeps, soft sweeps, clonal interference and the accumulation of drug resistance mutations. *G3: Genes/Genomes/Genetics*, 10(4):1213–1223.

Zayed, A. (2004). Effective population size in Hymenoptera with complementary sex determination. *Heredity*, 93(6):627–630.