Integrating Pool-seq uncertainties into demographic inference

João Carvalho [1], Hernán E. Morales [2], Rui Faria [3,4], Roger K. Butlin [5,6], and Vítor C. Sousa [1]

[1]cE3c - Centre for Ecology, Evolution and Environmental Changes & CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Portugal
[2]Section for Hologenomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark
[3]CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório Associado, Universidade do Porto, Vairão, Portugal
[4]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal
[5]Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Sheffield, S10 2TN United Kingdom
[6]Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden

Corresponding author's email address: jgcarvalho@fc.ul.pt

Abstract

Next-generation sequencing of pooled samples (Pool-seq) is a popular method to assess genome-wide diversity patterns in natural and experimental populations. However, Pool-seq is associated with specific sources of noise, such as unequal individual contributions. Consequently, using Pool-seq for the reconstruction of evolutionary history has remained underexplored. Here we describe a novel Approximate Bayesian Computation (ABC) method to infer demographic history, explicitly modeling Pool-seq sources of error. By jointly modeling Pool-seq data, demographic history and the effects of selection due to barrier loci, we obtain estimates of demographic history parameters accounting for technical errors associated with Pool-seq. Our ABC approach is computationally efficient as it relies on simulating subsets of loci (rather than the whole-genome), and on using relative summary statistics and relative model parameters. Our simulation study results indicate Pool-seq data allows distinction between general scenarios of ecotype formation (single versus parallel origin), and to infer relevant demographic parameters (e.g., effective sizes, split times). We exemplify the application of our method to Pool-seq data from the rocky-shore gastropod *Littorina saxatilis*, sampled on a narrow geographical scale at two Swedish locations where two ecotypes (Wave and Crab) are found. Our model choice and parameter estimates show that ecotypes formed before colonization of the two locations (i.e., single origin) and are maintained despite gene flow. These results indicate that demographic modeling and inference can be successful based on pool-sequencing using ABC, contributing to the development of suitable null models that allow for a better understanding of the genetic basis of divergent adaptation.

Keywords: Pool-seq, demographic inference, Approximate Bayesian Computation, R package, ecotype formation

Introduction

Population genomics data can be used to infer the complex demographic and adaptive processes that have shaped natural populations. Next Generation Sequencing (NGS) has revolutionized the field of population genomics, allowing reconstruction of evolutionary histories using thousands of SNPs across the genome (Ellegren, 2014). However, generating and sequencing individual libraries can be expensive and difficult for certain species (e.g., small organisms). In such cases, an effective alternative is to combine DNA from various individuals, producing a single library that is then sequenced (Pool-seq). NGS of pooled samples requires less DNA per individual, reducing the necessary laboratory work by decreasing the number of library preparations needed. This results in decreased costs while still allowing the comparison of populations on a genomic scale (Schlötterer, Tobler, Kofler, & Nolte, 2014). However, pooling introduces challenges in data analysis due to non-equimolar DNA concentrations and stochastic variation in amplification or sequencing efficiency, which can result in loss of accuracy of allele frequency estimates (Anderson, Skaug, & Barshis, 2014; Cutler & Jensen, 2010; Ellegren, 2014). Furthermore, DNA from multiple individuals can be extracted in batches, combining multiple batches into a single pool for library preparation and sequencing (Morales et al., 2019; Ross, Endersby-Harshman, & Hoffmann, 2019), which can lead to unequal representation due to variation in extraction efficiency and/or non-equimolar concentrations of DNA between batches. Nonetheless, theoretical and empirical comparisons of individual-based sequencing and Pool-seq indicate that when an equal sequencing effort is employed, Pool-seq allows the analysis of more individuals which leads to similar or more precise allele frequency estimates (Futschik & Schlötterer, 2010; Gautier et al., 2013). Although empirical studies showed that individual-based sequencing provides more information to detect fine-scale population substructure (e.g., hybrids and migrants) than Pool-seq, both approaches are suitable for inferring population genetic structure (Chen et al., 2022; Dorant et al., 2019). Indeed, when a large number of samples is available, Pool-seq data results in more accurate estimates of effective population sizes and divergence or admixture time events (Collin et al., 2021). Pool-seq has been used in various studies, ranging from population genomic analysis (Begun et al., 2007; Ferretti, Ramos-Onsins, & Pérez-Enciso, 2013; Rubin et al., 2012) to experimental evolution (Parts et al., 2011; Turner, Stewart, Fields, Rice, & Tarone, 2011; Zhou et al., 2011) and human genetics applications to uncover disease-related mutations (Calvo et al., 2010; Lieberman et al., 2014; Prescott et al., 2015). Yet, using Pool-seq to perform demographic history inference has been hampered by a lack of tools that explicitly model this type of data.

Recent developments in population genomics using simulations include machine learning (Schrider, Shanku, & Kern, 2018; Sheehan & Song, 2016) and model-based inference approaches. The latter allows comparing alternative models and estimating parameters. Model-based inference methods, such as Approximate Bayesian Computation (ABC), offer important advantages (for a review see Beaumont et al., 2010 and Hickerson, 2014), because they allow for explicit and joint consideration of evolutionary processes and sampling effects. ABC replaces data with summary statistics (e.g., heterozygosity, dxy, F_{ST}) and uses simulations to select models and estimate parameters. The simplest ABC algorithm is based on a rejection approach (Tavaré, Balding, Griffiths, & Donnelly, 1997), where parameter values (and/or models) sampled from the prior are accepted if the distance between the simulated and observed summary statistics is below a given distance threshold (i.e. tolerance) or rejected otherwise. Accepted parameter values provide a sample of independent points from the posterior distribution. Given its flexibility, ABC has been widely used in ecology (Pontarp, Brännström, & Petchey, 2019; Zhang, Dennis, Landers, Bell, & Perry, 2017), systems biology (Liepe et al., 2014) and population genetics (Cooke & Nakagome, 2018; Rougemont & Bernatchez, 2018), with various software implementations (Boitard, Rodríguez, Jay, Mona, & Austerlitz, 2016; Cornuet et al., 2014; Huang, Takebayashi, Qi, & Hickerson, 2011; Wegmann, Leuenberger, Neuenschwander, & Excoffier, 2010). However, implementing ABC for whole genome data is challenging (Smith & Flaxman, 2020) due to the heavy computational burden and difficulty in simulating recombination and mutation rate variation along the genome (Jay, Boitard, & Austerlitz, 2019).

Genomic data from natural populations has led to recent progress in the field of speciation, particularly through the study of ecotypes, which represent putative initial stages in speciation (Turesson, 1922). Many studies of ecotype evolution (Fang, Kemppainen, Momigliano, Feng, & Merilä, 2020; Ravinet et al., 2016; Riesch et al., 2017; Van Belleghem et al., 2018) aim to infer if the same phenotypes have evolved in multiple times and locations when facing similar divergent pressures, i.e. in parallel (Faria et al., 2014; Schluter, 2000). The support for natural selection in ecotype formation increases with the number of population replicates studied, but individual sequencing can become prohibitively expensive. Therefore, Pool-seq is useful in studies of parallel adaptation and speciation (Morales et al., 2019). Studies of ecotype formation usually consider two scenarios (Faria et al., 2014; Johannesson et al., 2010): (i) initial adaptive divergence occurred once with subsequent colonization of analogous pairs of environments (single origin scenario); and (ii) colonization of multiple environments was followed by independent evolutionary divergence (parallel origin scenario). Lower genetic distance between ecotypes within a locality, inferred by principal component analysis or structure plots, is frequently interpreted as a signal of parallel evolution. However, ongoing or past gene flow between different ecotypes can complicate the distinction between these scenarios (Faria et al., 2014). Rather, distinguishing between these hypotheses requires an explicit contrast of the different scenarios in a model-based framework (Butlin et al., 2012, 2014).

Model-based inference methods are commonly used to test whether divergence occurred with or without gene flow (Klütsch, Manseau, Trim, Polfus, & Wilson, 2016), whether there is ongoing gene flow (Bakovic et al., 2021), as well as in finding the most likely population tree for a given set of sampled populations (Louis et al., 2014) or estimating relevant demographic parameters (Andrew, Kane, Baute, Grassa, & Rieseberg, 2013). However, they have rarely been used explicitly to contrast different demographic scenarios of ecotype formation, despite some examples using coalescent-based approaches (Hume, Recknagel, Bean, Adams, & Mable, 2018) coupled with maximum composite-likelihoods (Le Moan, Gagnaire, & Bonhomme, 2016). Even in recognized model systems for parallel evolution in natural populations, such as the common rocky-shore gastropod, *Littorina saxatilis*, model-based inference methods have seldom been used. This species, found in locations that span the North Atlantic (Reid, 1996), is characterised by the existence of two ecotypes in close proximity: one adapted to crab predation (hereafter "Crab" ecotype) and another to heavier wave exposure ("Wave" ecotype) (Johannesson et al., 2010). Parallel differentiation of these ecotypes has been suggested before (Butlin et al., 2014; Panova, Hollander, & Johannesson, 2006; Rivas et al., 2018; Westram, Panova, Galindo, & Butlin, 2016) but only a single study, based on a limited number of markers, has contrasted the parallel origin scenario against an explicitly defined alternative hypothesis (Butlin et al., 2014). Thus, there is a clear need for efficient and easy-to-use methods that could readily distinguish between the two scenarios, particularly when that distinction might be complicated by recent gene flow.

Here we present a new R package to perform model choice and estimate demographic history parameters tailored to Pool-seq data. The main novelty is that we explicitly model and account for known sources of error associated with pool-based sequencing. We perform simulation studies to assess whether we can leverage pooled sequencing data to infer demographic parameters using ABC under a relatively simple two-population isolation with migration model and to differentiate between alternative scenarios of ecotype formation in more complex models with four populations. Importantly, we consider different migration rates among loci to account for the effects of selection against migrants at neutral markers linked to barriers against gene flow. We illustrate the application of our ABC method to Pool-seq whole genome data from *L. saxatilis* ecotypes, inferring whether the origin of the ecotypes consisted of a single or repeated parallel events in a narrow geographical area of two locations in Sweden.

Material and Methods

We developed an ABC method to model Pool-seq data explicitly under scenarios with two and four populations. Importantly, in all demographic models, we include an explicit parameter representing the error associated with the pooling process (e.g., unequal individual contribution) and a parameter representing errors associated with sequencing (e.g., sequencing and/or mapping errors). Below, we describe in detail the demographic models considered and the Pool-seq parameters in separate sub-sections.

Isolation with migration model with two populations

We started by considering a two population isolation with migration model with eight parameters (Figure 1A), assuming that an ancestral population of size N_{ref} (considered the reference effective size) splits T_{div} generations ago into two populations with constant effective population sizes N_1 and N_2 and with constant migration rates m_{12} and m_{21}. To account for the effects of linked selection due to barrier loci (i.e., effect of selection against migrants at neutral markers that are possibly linked to barriers against gene flow), we considered that a proportion of the genome P_{no} has no migration ($m_{12} = m_{21} = 0$).

Models with four populations: Single vs. parallel ecotype formation

To test the efficiency of our ABC method for distinguishing between different ecotype formation scenarios, we considered two alternative models with four populations. The four extant populations correspond to two ecotypes found at two different locations, i.e., two divergent ecotypes inhabiting each location. The four-population model has ten relevant demographic history parameters: the population size of the four extant populations (N_1 - N_4) and of the two ancestral populations (NA_1 and NA_2), the time of the recent (T_s) and ancient (T_{As}) split events in generations, and the two migration rates between the two populations ($m_{12} = m_{34}$, $m_{21} = m_{43}$) inhabiting each location. To estimate times of events, we considered as a parameter the time interval between the recent and the ancient split ($\Delta_s = T_{As} - T_s$). Migration rates between divergent ecotypes were assumed to be similar across the two geographic locations (e.g., $m_{12} = m_{34}$ - but note that the scaled migration rate $4Nm$ can be different). A proportion of loci (P_{no}) was also assumed to have no migration between different ecotypes. Depending on the topology, the four-population model can represent: (i) a single origin scenario, where ecotypes are formed in different locations, before dispersing to colonize the two geographic locations (panel B in Figure 1); or (ii) a parallel origin scenario, where colonization of each location is followed by independent and parallel divergence of the different ecotypes (panel C in Figure 1). Note that for the four-population models we assumed no migration between populations in different locations or between ancestral populations. Thus, the single origin model corresponds to a scenario of divergence of ecotypes without gene flow (i.e., no migration between ancestral populations), whereas in the parallel origin model the divergence of ecotypes occurs within each location with gene flow.

Coalescent simulations of individual genotypes

We used coalescent theory to simulate gene trees using *scrm* (Staab, Zhu, Metzler, & Lunter, 2015), under combinations of parameters and models sampled from the priors. Mutations were assumed to occur according to the infinite sites model, with a mutation rate μ per site and per generation. For each locus (i.e., window) in the genome we simulated gene trees with the same sample size, which corresponds to the number of individuals in the pool. In the simulation study we simulated pools of 100 diploid individuals (200 haplotypes) from each population. Thus, when simulating gene trees we assumed the actual haplotypes of all individuals in the pool were known and the effect of pooling was simulated at a later step (see next section). To simulate genotypes, we assumed that individuals within each population were reproducing at random and hence haplotypes were paired at random at each locus to obtain genotypes for each biallelic SNP.

Modelling Pool-seq data and combination of pools

To model allele frequencies at biallelic SNPs obtained with Pool-seq we follow a series of steps (Figure 2). Table 1 summarizes the notations used. Sample allele frequencies can be computed as the proportion of reads with a given allele. Thus, they are influenced by the depth of coverage at each single nucleotide polymorphism (SNP), which can vary along the genome due to NGS-associated stochasticity. To account for such variation, we considered that the number of reads at a given site follows a negative binomial distribution ($nBin$), previously shown to fit empirical distributions (e.g., Malaspinas et al. 2016). More precisely, we assumed that, for each SNP, the number of reads C_j for the j^{th} populations follows:

$$C_j \sim \ nBin(s, \psi) \tag{1}$$

where s and ψ are defined as:

$$s = \frac{mean(C_j)}{var(C_j)} \tag{2}$$

$$\psi = \frac{mean(C_j)^2}{var(C_j) - mean(C_j)} \tag{3}$$

where $mean(C_j)$ and $var(C_j)$ represent, respectively, the mean and variance of the depth of coverage across all SNPs of the j^{th} population. Another source of error in pool-based experiments is heterogeneity on the contribution of each individual to the DNA pool. PCR amplification step(s) during library preparation (e.g., for RAD markers; Baird et al. 2008) can also increase heterogeneity. To account for this uneven individual contribution we assumed that, for each site, the number of reads from the i^{th} individual ($r_{k,i}$) of the k^{th} pool follows a multinomial distribution:

$$r_{k,i} \sim \ mult(C_j, p_{k,i}) \tag{4}$$

where $p_{k,i}$ represents the expected proportion of reads from individual i in pool k, assumed to have a Dirichlet distribution:

$$p_{k,i} \sim \ Dir\left(\frac{\rho_i}{I_j}\right) \tag{5}$$

where I_j is the number of individuals of the j^{th} population, and ρ_i reflects the Pool-seq error (see below). When DNA extraction is performed for several pools of individuals that are combined into a larger pool, uneven contributions between pools might occur. To account for this variation, we assumed that the number of reads from the k^{th} pool (r_k) follows a multinomial distribution

$$r_k \sim \; mult(C_j, p_k) \tag{6}$$

where p_k is the expected proportion of reads from that pool, which follows a Dirichlet distribution:

$$p_k \sim \; Dir\left(\frac{\nu_{j,k}\rho_k}{I_j}\right) \tag{7}$$

where $\nu_{j,k}$ is the number of individuals in pool k of population j. Thus, our approach can be applied to pools with different sizes, ensuring that pools with more individuals have a higher contribution to the total number of reads. To obtain the number of reads for each individual inside each pool, we replaced C_j by r_k on equation 4, and I_j by $\nu_{j,k}$ on equation 5. Following Gautier et al. (2013), the unequal contributions of individuals and pools are modelled by increasing the variance of the proportion of reads, by adjusting ρ according to experimental error parameters ϵ_i and ϵ_p, for individuals ($p_{k,i}$) and pools (p_k), respectively. The corresponding variances are $var(p_{k,i}) = \left(\epsilon_i E[p_{k,i}]\right)^2$ and $var(p_k) = \left(\epsilon_p E[p_k]\right)^2$. The larger the experimental Pool-seq error (i.e. ϵ_i and ϵ_p), the larger the variance resulting in more unequal contributions from individuals. These can be used to derive ρ for individuals and pools (Gautier et al., 2013):

$$\rho_i = \left[\frac{\nu_{j,k}-1}{\nu_{j,k}^2 var(p_{k,i})}\right] - 1 = \left[\frac{\nu_{j,k}-1}{\nu_{j,k}^2 \left(\epsilon_i E[p_{k,i}]\right)^2}\right] - 1 \tag{8}$$

$$\rho_k = \left[\frac{K-1}{K^2 var(p_k)}\right] - 1 = \left[\frac{K-1}{K^2 \left(\epsilon_p E[p_k]\right)^2}\right] - 1 \tag{9}$$

where K is the total number of pools used to sequence the j^{th} population. In sum, this model ensures all individuals are expected to contribute the same number of reads, with errors due to unequal contribution modelled through the dispersion parameters ρ_i and ρ_k. When the experimental error rate tends to zero, the dispersion parameter tends to infinity, resulting in no pooling error as all individuals contribute exactly the same expected number of reads (Gautier et al., 2013). Finally, to account for sequencing and mapping errors, we assumed that, with an error rate ϵ_{seq}, ancestral allele A will be incorrectly called a derived allele D or vice-versa. More precisely, given the genotype and the total number of reads of the i^{th} individual at a given site, we assumed that the number of reads D_i with the derived allele follows a binomial distribution:

$$D_i \sim \begin{cases} Bin(r_{k,i}, \epsilon_{seq}) & \text{if individual is AA (homozygous ancestral)} \\ Bin(r_{k,i}, 1-\epsilon_{seq}) & \text{if individual is DD (homozygous derived)} \\ Bin(r_{k,i}, 0.5) & \text{if individual is AD (heterozygote)} \end{cases} \tag{10}$$

where $r_{k,i}$ represents the total number of reads contributed by a particular individual at a given site and ϵ_{seq} is the combined effect of both the sequencing and mapping errors. We assumed there are only two alleles at each site and that each base has an equal probability of being miscalled. Hence for heterozygotes each allele originates from either the ancestral or derived allele with equal probability (Li et al., 2012).

ABC implementation using subsets of loci

To avoid the computational burden of simulating whole genomes, we simulated sets of L independent loci with 2000 sites. We assumed that loci were independent, i.e., with free recombination between all pairs of loci ($r_b = 0.5$), and that within each locus of 2000 sites there was no recombination ($r_w = 0.0$). Our ABC implementation, based on a rejection algorithm, involved several steps: (i) sample demographic and pool-seq parameters from prior distributions (Table 2); (ii) simulate genotypes for each individual at L loci using coalescent gene trees based on demographic history parameters; (iii) simulate the number of reads and pooling of individuals for each biallelic SNP, applying filters (e.g., depth of coverage and minor allele frequency); (iv) compute summary statistics for observed and simulated data; (v) calculate Euclidean distance between observed and simulated summary statistics, standardizing to ensure that all summary statistics have the same mean and variance; (vi) reject parameters with distances above a tolerance threshold; (vii) apply a post-processing regression to adjust accepted parameter values (Beaumont, Zhang, & Balding, 2002). To simulate coalescent gene trees, we assumed all loci within a subset share the same demographic history, but set migration rate to zero at a proportion of loci P_{no} to account for selection effects due to barrier loci. For each resulting SNP, pool-seq data were simulated (Figure 2) by sampling depth of coverage from a negative binomial (equation 1) based on the observed mean and variance of the coverage of each population. To mimic common filter steps, we discarded SNPs with a depth of coverage outside a given range. For instance, for the *L. saxatilis* data, we kept only sites with a depth of coverage between 50x and 150x (see below). We then simulated each pool's contribution (equations 6 and 7) to the total coverage of a population and each individual's contribution to their pool's coverage (equations 4 and 5) by randomly sampling values from their respective distributions. Finally, we randomly drew the number of reads from the derived and ancestral alleles for each individual (equation 10), and then applied a filter to discard SNPs with fewer than two minor-allele reads. Note that we did not consider sequencing errors at invariant sites, as Pool-seq data were only simulated for polymorphic sites, and any such errors would be removed by the minor-allele frequency filter.

For each model, at least 5×10^5 simulations of $L = 300$ loci with $b = 2000$ base pairs were conducted. To reduce computational burden, parameter and summary statistic tables were saved and reused to analyze different subsets of loci from the observed data. To obtain posterior distributions, we combined 1000 subsets of $L = 300$ loci randomly selected from the observed data. Each subset was processed through steps (v) to (vii) of the ABC algorithm, resulting in a sample of independent points from the posterior of each parameter or model. We combined the independent posterior samples from the 1000 subsets of loci, taking into account the distance between the mean summary statistics of each subset and the overall mean across all loci in the genome. This was done using the Epanechnikov kernel, which assigns more weight to subsets of loci with means closer to the overall mean (supplementary Figure S1). Since demographic history is expected to affect all loci similarly across the genome, this approach aimed to minimize the impact of outlier subsets of loci on the posterior estimates. All steps were performed using custom-made functions and scripts in R, adapted from Beaumont et al. (2002).

Relative summary statistics and scaled parameters

We selected a set of statistics (Table S1) to summarize the patterns of relative diversity and differentiation within and among populations (Fraïsse et al., 2021; Jay et al., 2019), computed only for polymorphic sites across all populations. Namely, we considered: (i) expected heterozygosity per population and between all pairs of populations (Nei & Roychoudhury, 1974); (ii) pairwise F_{ST} between all pairs of populations (Bhatia, Patterson, Sankararaman, & Price, 2013); (iii) proportion of SNPs with fixed differences between populations (Fraïsse et al., 2021); (iv) proportion of exclusive SNPs within each population (Fraïsse et al., 2021); and for the four population models (v) several D-statistics with different combinations of P1, P2 and P3 populations (adapted from Malinsky, Matschiner, and Svardal (2021)). To capture the distribution across loci, we considered the mean and standard deviation of the above statistics. For F_{ST}, we further considered the 5% and the 95% quantiles because these should capture the effect of barriers to gene flow. In sum, we considered 13 summary statistics for the two-population model and 57 for the four-population models (Table S1).

Importantly, all these summary statistics are relative measures of diversity and differentiation that depend on relative branch lengths of coalescent trees (e.g., F_{ST}). Hence, we increased the efficiency of simulations by inferring relative demographic parameters scaled by the ancestral effective population size N_{ref}. We estimated relative effective sizes (e.g., $n_1 = N_1/N_{ref}$), relative times of divergence (e.g., $\delta_s = \Delta_s/4N_{ref}$), and scaled migration rates (e.g., $4N_1m_{21}$). To clarify, note that all relative parameters are represented with a lower case (e.g., n_1), while the absolute parameters are indicated with upper case letters (e.g., N_1) and that scaled migration rates specify which population is receiving immigrants by the subscript next to N. Estimation of relative parameters was done by performing coalescent simulations, fixing the ancestral effective population size to $N_{ref} = 25000$ and the mutation rate to $\mu = 1.5 \times 10^{-8}$ per site, as previously used for *L. saxatilis* (Butlin et al., 2014). To obtain absolute parameter estimates, we re-scaled parameters based on a re-scaling factor $f = obs[S]/E[S]$ that depends on the observed number of SNPs ($obs[S]$), and on the expected number of SNPs according to parameter estimates of a given model ($E[S]$). Assuming the infinite sites mutation model, the expected number of segregating sites was calculated based on the expected total branch length ($E[T]$), mutation rate per site (μ) and number of sites (L) as $E[S] = E[T]\mu L$ (Hudson, 1990). To obtain $E[T]$ we simulated 100,000 gene trees according to parameter estimates of a given model. The absolute effective population sizes and times of events in generations were obtained by multiplying by the rescaling factor f, i.e., $N_e = f \times n_e$ and $T_s = f \times t_s$, respectively.

Simulation study

For the two-population model, estimates were based on 10^6 simulations, whereas for the four-population scenarios they were based on 5×10^5 simulations for each scenario of ecotype formation. For each simulation, we generated 300 independent loci with 2000 base pairs, sampling 100 diploid individuals from each population. Parameter values were sampled from uniform or log-uniform prior distributions summarized in Table 2. The exception was the proportion without migration (P_{no}), which was sampled from a Beta distribution reflecting a low proportion of loci without migration *a priori*. For P_{no} we truncated the distribution, replacing values below 0.01 and above 0.50 by 0.00 and 0.50, respectively (Table 2). To evaluate the accuracy of our ABC implementation for Pool-Seq data to estimate parameters and model choice, we performed a leave-one-out cross-validation (Csilléry, François, & Blum, 2012). Hereafter, we use the term accuracy to indicate how close (or far off) a particular point estimate is to the true parameter value. Briefly, a random simulation was picked, and its summary statistics were used as pseudo-observed data. The remaining simulations were used to infer the parameters of the selected simulation. The ABC estimation was repeated for n pseudo-observed datasets. The prediction error was computed as:

$$\epsilon_{pred} = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n}(\hat{\Theta}_i - \Theta_i)^2}{var(\Theta)} \tag{11}$$

where Θ_i is the true parameter value of the i^{th} pseudo-observed dataset, $\hat{\Theta}_i$ is the estimated parameter value, and $var(\Theta)$ is the variance of the true parameter values. For parameter inference, we assessed the prediction error with $n = 5000$, considering three different point estimates (mode, median and mean of the posterior distribution), at two tolerance values (0.005 or 0.01). For comparison, we computed prediction errors using the mean of the prior distribution as point estimates. For evaluating the model choice we used $n = 1000$ pseudo-observed datasets. To define the model estimated for each pseudo-observed dataset, we considered two posterior probability thresholds: (i) 0.5, assigning a dataset to the model with posterior probability larger than 0.5; (ii) 0.9, a more stringent criterion assigning a dataset to a model only if the posterior was larger than 0.9, classifying it as ”unclear” otherwise.

Effect of explicitly modeling Pool-seq errors

By assuming that the proportion of reads with a given allele corresponds to the allele frequencies, it is possible to analyse Pool-seq data with existing model-based methods, e.g., fastsimcoal2 (Excoffier et al., 2021) and DIYABC random forest (DIYABC-RF) (Collin et al., 2021). Yet, ignoring Pool-seq associated errors due to unequal individual contribution might result in biased demographic estimates. To assess whether this is the case, and whether accounting for Pool-seq errors improves inference of demographic parameters, we compared estimates obtained either ignoring or explicitly modelling Pool-seq errors. We simulated a pseudo-observed Pool-seq dataset (i.e., with variable depth of coverage at each site and unequal individual contribution) according to the parameter estimates obtained for *L. saxatilis* with the two population model (Supplementary Table S7). We performed parameter inference using the regression adjustment with priors defined in Table 2 using $L = 100$ loci, 500k simulations and a tolerance of 0.01, either: (i) ignoring Pool-Seq errors by computing summary statistics directly from the simulated haplotypes; (ii) explicitly accounting for depth of coverage variation, unequal individual contribution and sequencing errors by computing summary statistics after simulating Pool-seq data as described above.

Effect of number of loci

To increase computational efficiency we simulate multiple subsets of loci, rather than entire genomes. To assess the impact of this strategy, we conducted 100k simulations with 10, 30, 100 or 300 simulated loci per subset using the two population isolation with migration model and the priors defined in Table 2. We then performed a leave-one-out cross-validation, as described above. We computed the prediction error using the mean of the regression-adjusted posterior as a point estimate for $n = 5000$ pseudo-observed datasets, with a tolerance of 0.01. To obtain the 95% confidence interval of the prediction error, we used a non-parametric bootstrap approach resampling 10k times the $n = 5000$ point estimates and re-calculating the prediction error.

Effect of combining multiple subsets of loci to obtain posteriors

Our method relies on combining posteriors obtained from multiple subsets of loci, giving more weight to subsets of loci with summary statistics closer to the mean whole-genome values. To evaluate the impact of this strategy we compared estimates obtained with the whole-genome with estimates obtained by merging the posteriors of random subsets of loci, varying the proportion of the genome sampled (10%, 30%, or 50% of the genome). To reduce the computational burden, we assumed that the whole-genome consisted of 100 loci. Using the two-population isolation with migration model, we generated 100 pseudo-observed whole-genomes according to the parameter estimates of *L. saxatilis.* Using the same model and the priors defined in Table 2, we conducted 100k simulations with 10, 30, or 50 loci per subset. Then, for each whole-genome, we sampled 100 subsets corresponding to either 10%, 30%, or 50% of the genome (i.e., 10, 30 or 50 loci). We performed parameter inference by merging the posteriors of the 100 subsets and using the regression adjustment with a tolerance of 0.01. These estimates were compared to the approach of Boitard et al. (2016), by using summary statistics computed from the whole-genome as a target to perform parameter inference, but using for inference summary statistics computed from a proportion of the genome, either with 10, 30 or 50 loci. We computed the bias of the estimates using $\frac{1}{n} \cdot \sum(\hat{\Theta}_i - \Theta_i)$, where $\hat{\Theta}_i$ is the estimated mean posterior with subsets of loci, and Θ_i is the mean posterior with 100 loci (mimicking the whole genome) for the i^{th} pseudo-observed dataset, while $n = 100$ is the number of simulated pseudo-observed datasets.

Impact of ignoring within-locus recombination

Our models assume free recombination between loci ($r_b = 0.5$) but no recombination within loci ($r_w = 0.0$). We evaluated the effect of this assumption on parameter estimates by comparing the posteriors obtained for pseudo-observed datasets with within-locus recombination to those obtained for datasets simulated without recombination to assess if ignoring within-locus recombination leads to changes in posteriors and thus impacts our estimates. This was done by simulating 100 pseudo-observed datasets according to the estimates obtained for *L. saxatilis* (see Supplementary Table S7). Each dataset contained 100 loci with within-locus recombination rate equal to the mutation rate ($r_w = \mu$). We then estimated the parameters using the regression adjustment with 500k simulations and a tolerance of 0.01, under our assumption of no within-locus recombination.

Littorina saxatilis **Pool-seq data**

We illustrate the application of our ABC implementation to previously published Pool-seq data (Morales et al., 2018) from *L. saxatilis* populations sampled at two different sites in Sweden (Arsklovet and Ramsö). At each of those sites, 100 females of the Crab and another 100 females of the Wave ecotype were sequenced in two separate pools (Morales et al., 2019). DNA extraction was performed for batches of five individuals by combining pieces of foot muscle tissue from five snails in one tube. Reads were trimmed with Trimmomatic v.0.36 (Bolger, Lohse, & Usadel, 2014) and mapped against the *L. saxatilis* reference genome, produced from a single Crab ecotype individual (Westram et al., 2018), using CLC v5.0.3 (`www.qiagenbioinformatics.com`). Only those reads with a mapping score higher than Q20 were retained. Bam files were processed with SAMtools v1.3.1 (Danecek et al., 2021), BEDtools v2.25.0 (Quinlan & Hall, 2010), and Picard tools v2.7.1 (`http://broadinstitute.github.io/picard`) and, for each set of bam files, reads with base quality lower than 30, mapping quality lower than 20 and those that mapped to very short contigs (<500 bp) were filtered out. We removed sites with a coverage lower than 50x or higher than 150x, ensuring we discarded low-coverage sites that would not contain reads for most individuals (i.e. <50x) and sites at putative repetitive or duplicated regions leading to unusually high depth of coverage (>150x). Recent studies have uncovered an important role of chromosomal inversions in the adaptive divergence of *L. saxatilis* ecotypes (Faria et al., 2019; Koch et al., 2021; Morales et al., 2019). Each inversion likely has its unique evolutionary history that may be influenced by various demographic and selective processes, such as divergent and balancing selection, and may differ from the population history. Therefore, to avoid biased estimates, inversion-tailored inference methods would be required, accounting for specific features such as varying recombination rates between homozygotes and heterozygotes. Since our aim was to infer the demographic history, an approach tailored to inversions is outside the scope of this study. Thus, we took a conservative approach removing regions that could be associated or linked with the reported inversions (Westram, Faria, Johannesson, & Butlin, 2021) (list of kept and removed contigs in Supplementary Data File 1). As breakpoints are not yet defined for many inversions, we removed 3671 contigs within inversions or in buffer regions. This corresponds to 3.3% of the whole genome Pool-seq dataset, distributed across the genome but with approximately 1/3 of the removed contigs located in chromosomes 10 and 12. To maximize the number of SNPs we kept all the remaining contigs, although only 20% of them map to known collinear regions (Westram et al., 2018). We estimated parameters of the two-population model for the two ecotypes from Arsklovet using the prior distributions and 10^6 simulations used for the simulation study. Similarly, we performed model choice and estimated parameters for the four-population models using 5×10^5 simulations, estimating parameters for the model with the highest posterior probability. Keeping in line with our strategy of using subsets of loci, we considered each contig in the *L. saxatilis* dataset as an independent locus and randomly selected a subset of contigs (i.e. $L = 300$). We then selected a random window of $b = 2000$ base pairs from each contig and computed summary statistics for those windows. To estimate parameters we computed the mean posterior (point estimate) and 95% credible intervals based on the weighted quantiles. Since this dataset only contained SNPs, remaining sites could be monomorphic or missing data. To re-scale the parameters, we calculated the number of SNPs per window assuming that the remaining sites were monomorphic. We converted time of events in generations to years, assuming a generation time of 0.5 years (Butlin et al., 2014).

Results

Performance of ABC point estimates

To evaluate the performance of our ABC implementation we performed a simulation study, summarizing the posterior distributions with three point estimates (mean, median and mode). When using $L = 300$ loci, prediction errors were lower using the mean or median with the regression-based adjustment for all the parameters (Tables S2, S3 and S4). As expected, with the regression, tolerance had a negligible effect in the prediction error. Additionally, prediction errors decreased with increasing number of simulated loci in the subsets, despite a clear trend of diminishing returns with more than $L = 100$ loci (supplementary Figure S2). Thus, unless specified, hereafter we summarize results obtained with subsets of $L = 300$ loci, using the regression-based adjustment and the mean as a point estimate, with a tolerance of 0.01.

Although the set of summary statistics was different for the two and four-population models, the prediction errors were similar for most parameters (supplementary Figures S3 and S4). For the relative effective sizes of extant populations (Figure 1), prediction errors ranged from 0.110 to 0.119 for the two-population model (Table 3, panel A in Figure 3), from 0.111 to 0.127 for the single origin (Figure 3B), and from 0.121 to 0.140 for the parallel origin (Table 3), indicating that the mean of posteriors provide accurate point estimates. For the sizes of ancestral populations in the four-population models (absolute values indicated by NA_1 and NA_2 in Figure 1), prediction errors were higher in the single origin than in the parallel origin (Table 3). For both models, the relative sizes of ancestral populations, na_1 and na_2, attained the highest prediction errors across all parameters, ranging from 0.530 to 0.616, indicating that point estimates are less accurate for ancestral effective sizes. Nevertheless, since prediction errors are smaller than the ones obtained when using the mean of the prior (close to 1), the shape of the posterior indicates that the summary statistics provide information about such parameters. For the relative timing of the split events, prediction errors were higher in the two-population model (0.34, Table 3, Figure 3D), than in the four-population models (ranging from 0.036 to 0.182). For the relative time of recent split (t_s), prediction errors were lower in the single origin model (0.036) than in the parallel model (0.172, Table 3), whereas for the relative time interval between split events (δ_s), prediction errors were similar for both models (0.182 for single, 0.179 for parallel) (Figure 3C-F and supplementary Figure S4 B-C, H-I).

Regarding the migration rates, although we specified prior immigration rates m_{ij} (probability that a lineage migrates from population i to j forward in time per generation), we focus on the average number of immigrants per generation ($4N_jm_{ij}$, where N_j is the effective size of the population receiving immigrants) as it accounts for both migration (proportional to m_{ij}) and drift (proportional to N_j), with $4N_jm_{ij} > 1$ indicating that migration occurs at a higher rate than drift. Prediction errors for $4N_jm_{ij}$ were similar for the two and four-population models, ranging from 0.284 to 0.340 (Table 3), although slightly higher in the parallel origin model. Across all models, the accuracy of the mean of the posterior decreased when the immigration used in simulations was too high, with poorer estimates when true $4N_jm_{ij} >> 10$. Overall, prediction errors for $4N_jm_{ij}$ were higher than for times of split and extant effective sizes, indicating that it is harder to accurately infer migration. The proportion of loci without migration (P_{no}) was accurately estimated, as supported by the very low prediction errors for the two and four-population models (Table 3).

Ignoring pooling and sequencing errors resulted in biased estimates for most demographic parameters (Figure 4 and supplementary Table S5), when pseudo-observed Pool-seq data were analysed without modelling explicitly the joint effect of variation in depth of coverage, unequal individual contribution and sequencing errors. Importantly, this is ignored by current demographic inference approaches (e.g., DIYABC-RF or fastsimcoal2). In contrast, our ABC approach based on explicitly modelling these sources of Pool-seq error provides accurate estimates (Figure 4). Although our aim was to demonstrate the implementation of an ABC method to perform parameter inference and model selection while explicitly modelling Pool-Seq data, treating pooling and sequencing errors as nuisance parameters, we report the prediction error for those parameters. The accuracy of the inference of the pooling error was similar to that of other parameters, with errors ranging from 0.241 to 0.243 (Table 3). This parameter was reasonably well estimated by the posterior mean when simulations were done with pooling errors above 150% (Figure 3I and supplementary Figure S3F, S4F-L). For the sequencing error, prediction error was higher for the two population (0.592) than for the four population models (0.042 - 0.062, Table 3), probably because there is more information in models with more individuals.

Performance of model choice

Results of the simulation study indicate that our ABC implementation allows a distinction between the single and parallel origin scenarios considered. Out of the 1,000 pseudo-observed datasets analysed under each model, using a 50% posterior probability threshold, the model was correctly inferred for 975 datasets of parallel origin (mean posterior probability of 0.952), and for 937 of single origin (mean posterior probability of 0.927, Figure 5A). When the model with the highest posterior was incorrect, its posterior probability was substantially lower (0.703 when parallel was inferred as single, and 0.755 when single was inferred as parallel). Using a more stringent threshold of 90% posterior probability, ABC still allowed to disentangle the two scenarios. The number of pseudo-observed datasets for which the model was correctly inferred was 877 for the parallel origin (one incorrectly assigned to the single model and 122 classified as unclear), and 854 for the single origin (12 incorrectly assigned to parallel and 134 classified as unclear (Figure 5B).

Application to *L. saxatilis* dataset: effect of merging subsets of loci and recombination

For simplicity, we discuss results after re-scaling relative parameters to absolute effective sizes and time of events in years, using k to indicate thousands (Table 4 but see Table S7 for the relative estimates). Re-scaling was performed after combining the posterior distributions from multiple subsets of loci, giving more weight to subsets of loci with mean summary statistics closer to the mean over the whole genome. By comparing posteriors obtained by merging subsets with varying numbers of loci, we found that using subsets of loci led to posteriors similar to those obtained with the whole genome (i.e. 100 simulated loci), but with a wider variance, i.e., higher uncertainty. Yet, even with subsets representing only 10% of the genome, posteriors were similar to those obtained using all loci, becoming closer as the number of loci in subsets increases (Figure 6, Supplementary table S6). Additionally, for all parameters, the bias obtained when merging posteriors is similar to, or lower than the bias obtained using the summary statistics from all SNPs to estimate parameters simulating just a subset of loci (as proposed by Boitard et al. (2016), supplementary table S6). Estimates based on the two-population model with Crab and Wave populations from Arsklovet indicate (Figure 7 and supplementary Figure S5): (i) a slightly larger effective size for Crab (mean ∼18k, 95% CI: 12k - 33k) than Wave (mean ∼15k, 95% CI: 10K - 28k) which, despite the large overlap of the CIs, is in line with previous studies using individual genotypes (a combination of mtDNA, amplified fragment length polymorphism markers and three nuclear genes) (Butlin et al., 2014); (ii) a split between Crab and Wave ecotype populations ∼18k years ago, but with a wide credible interval (95% CI: 2.2k - 111k); and that (iii) divergence was accompanied by gene flow, with higher immigration from the Crab into Wave ecotype, which is in agreement with reported cline shifts in these populations (Westram et al., 2021). Analysis of pseudo-observed datasets simulated under this scenario suggests that estimates are unlikely to be significantly biased by assuming no within-locus recombination ($r_w = 0$) since we obtained identical posterior distributions for pseudo-observed datasets simulated without ($r_w = 0$) or with a within-locus recombination rate equal to the mutation rate ($r_w = \mu$, supplementary Figure S6).

Our analysis of Crab and Wave ecotypes from two locations in Sweden (Arsklovet and Ramsö) supports the single origin model with strong posterior probabilities of 0.967 using the rejection algorithm and 1.000 using logistic regression. Our parameter estimates under the single origin model (Table 4 and supplementary Figure S7) suggest that the two ecotypes diverged approximately 15,000 years ago (95% CI: 5000 to 43000 years), followed by a recent colonization of both locations by populations from both ecotypes about ∼500 years ago (95% CI: 300 to 800 years). Under the single origin model, we estimated high and similar immigration rates between ecotypes in Arsklovet and lower migration from Wave into Crab in Ramsö (Figure 7H,I and Table S7). The point estimates supported larger ancestral effective sizes for the Crab population (mean 40k, 95% CI: 9K - 53k) than the Wave population (mean 21k, 95% CI: 4K - 48k), but the posteriors were wide and overlapping, indicating high uncertainty (Figure 7C). Nevertheless, the joint posteriors of present-day and ancestral populations indicate a population decline for the Crab ecotype in both locations, and for the Wave ecotype at Arsklovet. Finally, we inferred a proportion of loci without migration P_{no} close to zero, with a mean of approximately 1% and an upper CI close to 6% (Table S7).

Discussion

We developed a model-based method to analyse pooled-sequencing data, explicitly modeling various sources of error (e.g., variation in depth of coverage, unequal individual contribution, merging multiple pools) by extending the framework of Gautier et al. (2013) into an ABC inference framework. We implemented this into a freely available R package, allowing users to perform model choice and parameter inference of demographic history based on Pool-seq data from natural populations. Our approach is based on simulating subsets of loci, estimating relative parameters and using relative summary statistics. These included summary statistics that are widely used in ABC, such as the mean and standard deviation of expected heterozygosity per population and between all pairs of populations (Jay et al., 2019), relative genetic differentiation between population pairs (F_{ST}), and others that capture parts of the joint site frequency spectrum (Wakeley & Hey, 1997), such as the proportion of SNPs with fixed difference between populations (Fraïsse et al., 2021). To increase computational efficiency we fixed the ancestral effective population size (N_{ref}) and inferred relative demographic parameters, which were converted to absolute values based on an average mutation rate and number of observed SNPs. This circumvented the simulation of combinations of parameters leading to similar diversity and differentiation values, e.g., identical $\theta = 4N_e\mu$ and hence identical summary statistics due to low N_e with high μ or high N_e with low μ. Moreover, by combining multiple posterior distributions, obtained from different subsets of independent loci, and weighting them according to the distance to the genome-wide mean summary statistics, we minimized the impact of non-neutral processes (e.g., background selection) in the inference of demographic history. Our simulation study shows that, for the datasets analyzed here, the means of the posterior distributions provide accurate point estimates for most demographic history parameters of the two- and four-population models. In fact, the prediction errors for most parameters were similar for both models (Table 3), with the exception of migration rates, for which we found higher prediction errors for the parallel origin model (Table 3). This can be explained by the recent divergence of ecotypes with gene flow in each location, implying that it is harder to disentangle gene flow from incomplete lineage sorting under the parallel origin model. Importantly, our prediction errors based on Pool-seq were within the range of those of recent ABC methods based on individual genotypes (Fraïsse et al., 2021). Although the aim was to infer demographic history accounting for the effects of barrier loci, results indicate that the proportion of loci without migration (P_{no}) was well estimated in the two- and four-population models, suggesting it is possible to estimate the number of barrier loci under selection. Additionally, and despite concerns about model choice and estimation of Bayes factors with ABC (Marin, Pillai, Robert, & Rousseau, 2014; Robert, Cornuet, Marin, & Pillai, 2011), our model choice results indicate that Pool-seq provides enough information to distinguish between scenarios of ecotype formation with high posterior probabilities (proportion of correctly assigned simulations with 90% posterior probability above 0.85 for both models, Figure 5). This is explained by the fact that the single and parallel origin models considered have different mean values for several summary statistics (supplementary Figure S8), which is required to distinguish models in an ABC framework (Marin et al., 2014), and was expected given that gene flow occurs between populations with different shared ancestries in the alternative models (Figure 1). Importantly, our R package includes functions to compute prediction errors, allowing users to perform simulation studies based on their specific set of models, prior distributions, sample sizes, depths of coverage and numbers of pools. Thus, users can evaluate the accuracy of ABC results for their specific datasets and models. Also, the R package includes functions to assess the fit of the models to the data, visually plotting the fit of simulations to the observed summary statistics. Below we discuss the application to *L. saxatilis* ecotypes, as well as limitations and future perspectives.

Recent single origin of *Littorina saxatilis* ecotypes in Sweden

To illustrate the application of our method to Pool-Seq data, we analysed data from pools of *L. saxatilis* ecotypes, exploring the effects of obtaining posteriors by merging subsets of loci and assumptions about within-locus recombination. Using subsets of 300 loci, we found evidence supporting a single origin of Crab and Wave ecotypes in Sweden. Our results indicate that the ecotypes diverged relatively recently, followed by a split of the populations in different locations about 1,000 generations ago (approximately 500 years ago), with high gene flow between ecotypes. This is consistent with a recent postglacial colonization of Swedish islands (Panova et al., 2011). The estimates from both the two- and four-population models were consistent, with the divergence time for Crab and Wave ecotypes being approximately 15,000 years ago. Both models also indicate high migration rates between ecotypes ($4Nm > 10$), with slightly higher rates from Crab to Wave ecotypes (Figure 7G-I). This supports the hypothesis of a higher net dispersal from Crab to Wave, which may explain the observed shift in cline centres towards the wave habitat on Swedish islands (Westram et al., 2021). We found slightly larger effective sizes for Crab than Wave ecotype populations, together with lower effective sizes for present-day than ancestral populations, in agreement with a previously reported lack of support for past expansions based on individual genotypes (Butlin et al., 2014). Despite the high uncertainty in the posteriors for ancestral population sizes, our estimates suggest a higher density of individuals in Crab than Wave habitats, which is also consistent with the reported shifts in cline centres (Westram et al., 2021). Finally, we found that a low proportion of the genome was linked to complete barriers to gene flow between the two ecotypes ($P_{no} < 6\%$). This low proportion of barrier loci was not surprising since we excluded SNPs from all known regions associated with chromosomal inversions in *L. saxatilis*, which play an important role in the non-neutral ecotype divergence process (Westram et al., 2021). Thus, a possible explanation for our estimates is that barrier loci also occur outside inversions. However, given the lack of a chromosome level reference genome with a clear mapping of collinear and inverted regions, we cannot exclude that some of the SNPs included in our analysis are actually linked with chromosomal inversions.

The inferred high gene flow ($4Nm > 10$) between Wave and Crab populations may limit our ability to distinguish between alternative models (Bierne, Gagnaire, & David, 2013), but our results and ABC model choice based on individual genotypes (Butlin et al., 2014) both support a single origin for *L. saxatilis* ecotypes in Sweden. Indeed, simulations under the single origin model fit the observed summary statistics (supplementary Figure S9), but caution is needed due to the simplified nature of our models. Due to the limited spatial scale of our study, our results may reflect recent postglacial colonization of the two locations, rather than ecotype formation. Indeed, it is probable that ecotype formation in these Swedish locations predates their colonization. To determine if ecotype formation occurred in parallel, the ABC approach developed here could be applied to compare Wave and Crab ecotypes from more distant locations.

Limitations and future perspectives

Our aim was to implement an ABC method using Pool-seq data and test its performance under generic two- and four-population divergence models. These models are relatively simple, and probably fail to capture the complexity of ecotype formation in these geographically restricted *L. saxatilis* Crab and Wave ecotypes. For instance, we assumed a simultaneous divergence of the four extant populations, and no migration between ancestral populations, which is unlikely to hold. More complex models, implying different strengths of selection at barrier loci or the possibility of one ecotype acting as a reservoir of standing genetic variation (Jones et al., 2012; Liu, Ferchaud, Grønkjær, Nygaard, & Hansen, 2018) could also be considered. It remains to be tested whether an ABC framework allows distinguishing between more complex models with Pool-seq data. Nonetheless, a recent study has highlighted the potential of Pool-seq data to infer demographic histories by combining ABC with supervised machine learning in the DIYABC-RF software (Collin et al., 2021). Similarly to our approach, DIYABC-RF enables the simulation and analysis of Pool-seq data by first simulating individual SNP genotypes and then using the corresponding allele frequencies to generate pool read counts from a binomial distribution. However, DIYABC-RF does not explicitly model all possible sources of Pool-seq errors, as it only models variation in read coverage across SNPs (by randomly drawing coverages from the vectors of SNP coverages in the observed data set). Here, we explicitly model the different sources of errors with specific error parameters, such as variation in depth of coverage, unequal individual and pool contributions, and sequencing errors. Our results show that ignoring Pool-Seq errors might lead to incorrect estimates, but that demographic parameters are estimated accurately by explicitly modeling Pool-Seq errors (Figure 4). The low prediction errors found in our simulation study in models with up to four populations indicate that Pool-seq data might be suitable to infer demographic history under more complex models.

Our modular approach allows users to integrate our R package seamlessly with other packages at different steps. First, here we used the coalescent simulator implemented in the R package *scrm*, but it is possible to consider other demographic scenarios and simulate genetic data with coalescent-based methods for recombining chromosomes (Kelleher, Etheridge, & McVean, 2016), or forward simulators that explicitly model positive and background selection (Haller & Messer, 2019) and then use our functions to simulate Pool-seq data. Second, after simulating Pool-seq data, users can feed the reference tables with parameters and summary statistics to other tools using more sophisticate algorithms, such as neural networks or random forest ABC. Third, after the ABC rejection step, users can perform post-processing adjustment using other tools (e.g., abc R package, Csilléry et al. 2012). Despite some limitations, our results show that combining Pool-seq with ABC is an effective approach for investigating parallel evolution in taxa where similar ecotypes are found at multiple locations. We illustrated this by applying our method to Swedish populations of *L. saxatilis* ecotypes. The demographic history models considered provide suitable null models for a better comprehension of the genetic basis of divergent adaptation across many taxa.

Acknowledgements

This work was funded by the strategic project UIDB/00329/2020 granted to cE3c from the Portuguese National Science Foundation — Fundação para a Ciência e a Tecnologia (FCT). JC was supported by an FCT Ph.D. scholarship (PD/BD/128350/2017). RF is funded by a FCT CEEC (Fundação para a Ciênca e a Tecnologia, Concurso Estímulo ao Emprego Científico) contract (2020.00275.CEECIND) and by a FCT research project (PTDC/BIA-EVL/1614/2021). RKB was funded by the European Research Council (ERC-2015-AdG-693030-BARRIERS). VCS was supported by Fundação Ciência e Tecnologia (CEECINST/00032/2018/CP1523/CT0008) and by Human Frontier Science Program (RGY0081/2020). We thank the National Network for Advanced Computing (RNCA) and INCD (https://incd.pt/) for access and use of their computing infrastructure, funded by FCT to VCS (2021.09795.CPCA).

References

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502-512. doi: 10.1111/mec.12609

Andrew, R. L., Kane, N. C., Baute, G. J., Grassa, C. J., & Rieseberg, L. H. (2013). Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, *22*(3), 799-813. doi: 10.1111/mec.12038

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid snp discovery and genetic mapping using sequenced rad markers. *PLOS ONE*, *3*(10), 1-7. doi: 10.1371/journal.pone.0003376

Bakovic, V., Martin Cerezo, M. L., Höglund, A., Fogelholm, J., Henriksen, R., Hargeby, A., & Wright, D. (2021). The genomics of phenotypically differentiated *Asellus aquaticus* cave, surface stream and lake ecotypes. *Molecular Ecology*, *30*(14), 3530-3547. doi: 10.1111/mec.15987

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., ... Excoffier, L. (2010). In defence of model-based inference in phylogeography. *Molecular Ecology*, *19*(3), 436-446. doi: 10.1111/j.1365-294X.2009.04515.x

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, *162*(4), 2025-2035. doi: 10.1111/j.1937-2817.2010.tb01236.x

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., ... others (2007). Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans. *PLoS biology*, *5*(11), e310. doi: 10.1371/journal.pbio.0050310

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting fst: the impact of rare variants. *Genome research*, *23*(9), 1514-1521. doi: 10.1101/gr.154831.113.

Bierne, N., Gagnaire, P.-A., & David, P. (2013). The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology*, *59*(1), 72-86. doi: 10.1093/czoolo/59.1.72

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. *PLOS Genetics*, *12*(3), 1-36. doi: 10.1371/journal.pgen.1005877

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Butlin, R. K., Debelle, A., Kerth, C., Snook, R. R., Beukeboom, L. W., Cajas, R. C., … others (2012). What do we need to know about speciation? *Trends in Ecology & Evolution*, *27*(1), 27-39. doi: 10.1016/j.tree.2011.09.002

Butlin, R. K., Saura, M., Charrier, G., Jackson, B., André, C., Caballero, A., … Rolán-Alvarez, E. (2014). Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, *68*(4), 935-949. doi: 10.1111/evo.12329

Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., … Mootha, V. K. (2010). High-throughput, pooled sequencing identifies mutations in nubpl and foxred1 in human complex i deficiency. *Nature Genetics*, *42*(10), 851-858. doi: 10.1038/ng.659

Chen, C., Parejo, M., Momeni, J., Langa, J., Nielsen, R. O., Shi, W., … others (2022). Population structure and diversity in european honey bees (apis mellifera l.)—an empirical comparison of pool and individual whole-genome sequencing. *Genes*, *13*(2), 182. doi: 10.3390/genes13020182

Collin, F.-d., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., … Estoup, A. (2021). Extending approximate bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using diyabc random forest. *Molecular Ecology Resources*, *21*(8), 2598-2613. doi: 10.1111/1755-0998.13413

Cooke, N. P., & Nakagome, S. (2018). Fine-tuning of approximate bayesian computation for human population genomics. *Current Opinion in Genetics and Development*, *53*, 60-69. doi: 10.1016/j.gde.2018.06.016

Cornuet, J. M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., … Estoup, A. (2014). Diyabc v2.0: A software to make approximate bayesian computation inferences about population history using single nucleotide polymorphism, dna sequence and microsatellite data. *Bioinformatics*, *30*(8), 1187-1189. doi: 10.1093/bioinformatics/btt763

Csilléry, K., François, O., & Blum, M. G. (2012). Abc: An r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, *3*(3), 475-479. doi: 10.1111/j.2041-210X.2011.00179.x

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41-43. doi: 10.1534/genetics.110.121012

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of samtools and bcftools. *Gigascience*, *10*(2), giab008. doi: 10.1093/gigascience/giab008

Dorant, Y., Benestan, L., Rougemont, Q., Normandeau, E., Boyle, B., Rochette, R., & Bernatchez, L. (2019). Comparing pool-seq, rapture, and gbs genotyping for inferring weak population structure: The american lobster (homarus americanus) as a case study. *Ecology and evolution*, *9*(11), 6606-6623. doi: 10.1002/ece3.5240

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51-63. doi: 10.1016/j.tree.2013.09.008

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882-4885. doi: 10.1093/bioinformatics/btab468

Fang, B., Kemppainen, P., Momigliano, P., Feng, X., & Merilä, J. (2020). On the causes of geographically heterogeneous parallel evolution in sticklebacks. *Nature ecology & evolution*, *4*(8), 1105-1115. doi: 10.1038/s41559-020-1222-6

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., ... Butlin, R. K. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, *28*(6), 1375-1393. doi: 10.1111/mec.14972

Faria, R., Renaut, S., Galindo, J., Pinho, C., Melo-Ferreira, J., Melo, M., ... Butlin, R. K. (2014). Advances in ecological speciation: an integrative approach. *Molecular Ecology*, *23*(3), 513-521. doi: 10.1111/mec.12616

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561-5576. doi: 10.1111/mec.12522

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., ... Roux, C. (2021). Dils: Demographic inferences with linked selection by using abc. *Molecular Ecology Resources*, *21*(8), 2629-2644. doi: 10.1111/1755-0998.13323

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled dna samples. *Genetics*, *186*(1), 207-218. doi: 10.1534/genetics.110.114397

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., ... Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766-3779. doi: 10.1111/mec.12360

Haller, B. C., & Messer, P. W. (2019). Slim 3: forward genetic simulations beyond the wright-fisher model. *Molecular biology and evolution*, *36*(3), 632-637. doi: 10.1093/molbev/msy228

Hickerson, M. J. (2014). All models are wrong. *Molecular Ecology*, *23*(12), 2887-2889. doi: 10.1111/mec.12794

Huang, W., Takebayashi, N., Qi, Y., & Hickerson, M. J. (2011). Mtml-msbayes: Approximate bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. *BMC Bioinformatics*, *12*(1), 1-14. doi: 10.1186/1471-2105-12-1

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, *7*(1), 1-44.

Hume, J. B., Recknagel, H., Bean, C. W., Adams, C. E., & Mable, B. K. (2018). Radseq and mate choice assays reveal unidirectional gene flow among three lamprey ecotypes despite weak assortative mating: insights into the formation and stability of multiple ecotypes in sympatry. *Molecular ecology*, *27*(22), 4572-4590. doi: 10.1111/mec.14881

Jay, F., Boitard, S., & Austerlitz, F. (2019). An abc method for whole-genome sequence data: inferring paleolithic and neolithic human expansions. *Molecular biology and evolution*, *36*(7), 1565-1579. doi: 10.1093/molbev/msz038

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., & Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1547), 1735-1747. doi: 10.1098/rstb.2009.0256

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Team, B. I. G. S. P. . W. G. A. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55-61. doi: 10.1038/nature10944

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, *12*(5), e1004842. doi: 10.1371/journal.pcbi.1004842

Klütsch, C. F. C., Manseau, M., Trim, V., Polfus, J., & Wilson, P. J. (2016). The eastern migratory caribou: the role of genetic introgression in ecotype evolution. *Royal Society Open Science*, *3*(2), 150469. doi: 10.1098/rsos.150469

Koch, E. L., Morales, H. E., Larsson, J., Westram, A. M., Faria, R., Lemmon, A. R., … Butlin, R. K. (2021). Genetic variation for adaptive traits is associated with polymorphic inversions in littorina saxatilis. *Evolution Letters*, *5*(3), 196-213. doi: 10.1002/evl3.227

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). Popoolation2: identifying differentiation between populations using sequencing of pooled dna samples (pool-seq). *Bioinformatics*, *27*(24), 3435-3436. doi: 10.1093/bioinformatics/btr589

Le Moan, A., Gagnaire, P.-A., & Bonhomme, F. (2016). Parallel genetic divergence among coastal-marine ecotype pairs of european anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187-3202. doi: 10.1111/mec.13627

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., … Abecasis, G. R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. *PLOS Genetics*, *8*(10), 1-12. doi: 10.1371/journal.pgen.1002944

Lieberman, T. D., Flett, K. B., Yelin, I., Martin, T. R., McAdam, A. J., Priebe, G. P., & Kishony, R. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature Genetics*, *46*(1), 82-87. doi: 10.1038/ng.2848

Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., & Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature Protocols*, *9*(2), 439-456. doi: 10.1038/nprot.2014.025

Liu, S., Ferchaud, A.-L., Grønkjær, P., Nygaard, R., & Hansen, M. M. (2018). Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. *Molecular ecology*, *27*(23), 4725-4743. doi: 10.1111/mec.14782

Louis, M., Fontaine, M. C., Spitz, J., Schlund, E., Dabin, W., Deaville, R., … Simon-Bouhet, B. (2014). Ecological opportunities and specializations shaped genetic divergence in a highly mobile marine top predator. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1795), 20141558. doi: 10.1098/rspb.2014.1558

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., … others (2016). A genomic history of aboriginal australia. *Nature*, *538*(7624), 207-214. doi: 10.1038/nature18299

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite-fast d-statistics and related admixture evidence from vcf files. *Molecular ecology resources*, *21*(2), 584-595. doi: 10.1111/1755-0998.13265

Marin, J.-M., Pillai, N. S., Robert, C. P., & Rousseau, J. (2014). Relevant statistics for bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(5), 833-859. doi: 10.1111/rssb.12056

Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2018). *Littorina saxatilis genome sequencing and population re-sequencing.* Retrieved from `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA494650`

Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: beyond a single environmental contrast. *Science advances*, *5*(12), eaav9963. doi: 10.1126/sciadv.aav9963

Nei, M., & Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics*, *76*(2), 379-390. doi: 10.1093/genetics/76.2.379

Panova, M., Blakeslee, A. M., Miller, A. W., Mäkinen, T., Ruiz, G. M., Johannesson, K., & André, C. (2011). Glacial history of the north atlantic marine snail, *Littorina saxatilis*, inferred from distribution of mitochondrial dna lineages. *PLoS One*, *6*(3), e17511. doi: 10.1371/journal.pone.0017511

Panova, M., Hollander, J., & Johannesson, K. (2006). Site-specific genetic divergence in parallel hybrid zones suggests nonallopatric evolution of reproductive barriers. *Molecular Ecology*, *15*(13), 4021-4031. doi: 10.1111/j.1365-294X.2006.03067.x

Parts, L., Cubillos, F. A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S. J., ... Liti, G. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, *21*(7), 1131-1138. doi: 10.1101/gr.116731.110

Pontarp, M., Brännström, Å., & Petchey, O. L. (2019). Inferring community assembly processes from macroscopic patterns using dynamic eco-evolutionary models and approximate bayesian computation (abc). *Methods in Ecology and Evolution*, *10*(4), 450-460. doi: 10.1111/2041-210X.13129

Prescott, N. J., Lehne, B., Stone, K., Lee, J. C., Taylor, K., Knight, J., ... Consortium, U. I. G. (2015). Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in btnl2 and implicates other immune related genes. *PLOS Genetics*, *11*(2), 1-19. doi: 10.1371/journal.pgen.1004955

Quinlan, A. R., & Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. doi: 10.1093/bioinformatics/btq033

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., & Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of littorina saxatilis at a local scale. *Molecular ecology*, *25*(1), 287-305. doi: 10.1111/mec.13332

Reid, D. G. (1996). *Systematics and evolution of littorina*. London: Ray Society.

Riesch, R., Muschick, M., Lindtke, D., Villoutreix, R., Comeault, A. A., Farkas, T. E., ... others (2017). Transitions between phases of genomic differentiation during stick-insect speciation. *Nature ecology & evolution*, *1*(4), 1-13. doi: 10.1038/s41559-017-0082

Rivas, M. J., Saura, M., Pérez-Figueroa, A., Panova, M., Johansson, T., André, C., ... Quesada, H. (2018). Population genomics of parallel evolution in gene expression and gene sequence during ecological adaptation. *Scientific reports*, *8*(1), 1-12. doi: 10.1038/s41598-018-33897-8

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences*, *108*(37), 15112-15117. doi: 10.1073/pnas.1102900108

Ross, P. A., Endersby-Harshman, N. M., & Hoffmann, A. A. (2019). A comprehensive assessment of inbreeding and laboratory adaptation in *Aedes aegypti* mosquitoes. *Evolutionary applications*, *12*(3), 572-586. doi: doi.org/10.1111/eva.12740

Rougemont, Q., & Bernatchez, L. (2018). The demographic history of atlantic salmon (salmo salar) across its distribution range reconstructed from approximate bayesian computations*. *Evolution*, *72*(6), 1261-1277. doi: 10.1111/evo.13486

Rubin, C.-J., Megens, H.-J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., … Andersson, L. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences*, *109*(48), 19529-19536. doi: 10.1073/pnas.1217149109

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749-763. doi: 10.1038/nrg3803

Schluter, D. (2000). *The ecology of adaptive radiation: Oxford university press.* Oxford University Press.

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS genetics*, *14*(4), e1007341. doi: 10.1371/journal.pgen.1007341

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS computational biology*, *12*(3), e1004845. doi: 10.1371/journal.pcbi.1004845

Smith, C. C., & Flaxman, S. M. (2020). Leveraging whole genome sequencing data for demographic inference with approximate bayesian computation. *Molecular ecology resources*, *20*(1), 125-139. doi: 10.1111/1755-0998.13092

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680-1682. doi: 10.1093/bioinformatics/btu861

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, *145*(2), 505-518.

Turesson, G. (1922). The genotypical response of the plant species to the habitat. *Hereditas*, *3*(3), 211-350.

Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., & Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in drosophila melanogaster. *PLOS Genetics*, *7*(3), 1-10. doi: 10.1371/journal.pgen.1001336

Van Belleghem, S. M., Vangestel, C., De Wolf, K., De Corte, Z., Möst, M., Rastas, P., . . . Hendrickx, F. (2018). Evolution at two time frames: polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS genetics*, *14*(11), e1007796. doi: 10.1371/journal.pgen.1007796

Wakeley, J., & Hey, J. (1997). Estimating ancestral population parameters. *Genetics*, *145*(3), 847-855. doi: 10.1093/genetics/145.3.847

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). Abctoolbox: a versatile toolkit for approximate bayesian computations. *BMC Bioinformatics*, *11*(1), 1-7. doi: 10.1186/1471-2105-11-116

Westram, A. M., Faria, R., Johannesson, K., & Butlin, R. K. (2021). Using replicate hybrid zones to understand the genomic basis of adaptive divergence. *Molecular ecology*, *30*(15), 3797-3814. doi: 10.1111/mec.15861

Westram, A. M., Panova, M., Galindo, J., & Butlin, R. K. (2016). Targeted resequencing reveals geographical patterns of differentiation for loci implicated in parallel evolution. *Molecular ecology*, *25*(13), 3169-3186. doi: 10.1111/mec.13640

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., . . . Butlin, R. K. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution letters*, *2*(4), 297-309. doi: 10.1002/evl3.74

Zhang, J., Dennis, T. E., Landers, T. J., Bell, E., & Perry, G. L. (2017). Linking individual-based and statistical inferential models in movement ecology: A case study with black petrels (procellaria parkinsoni). *Ecological Modelling*, *360*, 425-436. doi: 10.1016/j.ecolmodel.2017.07.017

Zhou, D., Udpa, N., Gersten, M., Visk, D. W., Bashir, A., Xue, J., . . . Haddad, G. G. (2011). Experimental selection of hypoxia-tolerant drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *108*(6), 2349-2354. doi: 10.1073/pnas.1010643108

Data Accessibility Statement

All custom scripts used to perform the simulations and Approximate Bayesian Analysis can be found in the GitHub repository: `https://github.com/joao-mcarvalho/poolABC`. These scripts will be made available as an R package on the CRAN repository upon publication. Genomic data from *Littorina saxatilis* populations was previously processed in Morales et al. (2019). All the custom scripts used by the authors can be found in the GitHub repository: `https://github.com/hmoral/Ls_pool_seq`. Raw sequencing reads were deposited in the Sequence Read Archive under the BioProject PRJNA494650. Additional data related to this paper may be requested from the authors.

Author Contributions

JC and VCS developed the theoretical formalism, performed the analytic calculations, and planned the study. JC developed and implemented the R package, and performed the simulation study to validate the inference method. HM processed the observed genomic data. JC and VCS analyzed the data. JC wrote the manuscript together with VCS, with support from RF and RKB. RF, RKB and VCS supervised the project. All authors provided critical feedback and helped shape the analysis and manuscript.

Table 1: **Summary of main notations used.** Note that when we refer to individuals throughout this table, we are referring to diploid individuals.

Notation	Parameter definition
l	Total number of populations in the pooling experiment
C_j	Total number of reads of the j^{th} population (total coverage)
K	Total number of pools used to sequence the j^{th} population
$\nu_{j,k}$	Total number of individuals sequenced in the k^{th} pool of the j^{th} population
$I_j = \sum\limits_{k=1}^{K} \nu_{j,k}$	Total number of individuals of population j
$I = \sum\limits_{j=1}^{l} \sum\limits_{k=1}^{K} \nu_{j,k}$	Total number of individuals in the pooling experiment
$E[p_k]$	Expected value of the contribution of the k^{th} pool
$E[p_{k,i}]$	Expected value of the contribution of i^{th} individual of the k^{th} pool
ρ	Pool-seq error, proportional to dispersion of individual (or pool) contribution around their expected value
p_k	Contribution (proportion) of reads from the k^{th} pool $\left(\sum\limits_{k=1}^{K} p_k = 1\right)$
$p_{k,i}$	Contribution (proportion) of reads from the i^{th} individual of the k^{th} pool of population j $\left(\sum\limits_{i=1}^{\nu_{j,k}} p_{k,i} = 1\right)$
r_k	Number of reads from the k^{th} pool $\left(r_k = \sum\limits_{i=1}^{\nu_{j,k}} r_{k,i}\right)$ of population j (pool coverage). Note that $C_j = \sum\limits_{k=1}^{K} r_k = \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{\nu_{j,k}} r_{k,i}$
$r_{k,i}$	Number of reads from the i^{th} individual of the k^{th} pool of a given population
D_i	Number of derived allele reads of the i^{th} individual

Table 2: **Prior distributions and their ranges for each parameter.** Parameters are presented for the four-population models and, when relevant, for the two-population model. n_i - relative sizes of the extant populations (n_1, n_2, n_3, n_4); na_i - relative sizes of the ancestral populations (na_1, na_2); t_{div} - relative time of the split event in the two-population model, t_s - relative time of the recent split event; δ_s - relative time interval between t_s and the ancient split event (t_{As}); ϵ_{pool} - experimental error introduced by the pooling procedures; ϵ_{seq} - error associated with sequencing and mapping errors; m_{ij} - probability per generation that an individual migrates from the N_1 or N_3 (Crab) population to the N_2 or N_4 (Wave) population (forward in time), m_{ji} - probability per generation that an individual migrates from the N_2 or N_4 (Wave) population to the N_1 or N_3 (Crab) population (forward in time) and P_{no} - proportion of the simulated loci where no migration occurs between ecotypes.

	Two-population model		Four-population models	
parameter	minimum	maximum	minimum	maximum
n_i	0.1	3	0.1	3
na_i	-	-	0.1	3
t_{div}	0	3	-	-
t_s	-	-	0	3
δ_s	-	-	0	3
ϵ_{pool}	5	250	5	250
ϵ_{seq}	0.0001	0.001	0.0001	0.001
m_{ij}	10^{-13}	10^{-3}	10^{-13}	10^{-3}
m_{ji}	10^{-13}	10^{-3}	10^{-13}	10^{-3}
P_{no}	0	0.5	0	0.5

Table 3: **Prediction errors for parameter estimation.** Prediction errors were computed using the mean of the posterior distribution, obtained after the regression adjustment and a tolerance of 0.01. Prior mean indicates the prediction error if the mean of the prior distribution were used as point estimates. n_1 to n_4 - relative population sizes of the extant populations; na_1 and na_2 - relative population sizes of the ancestral populations; t_{div} - relative time of the split event in the two-population model; t_s - relative time of the split event that lead to the origin of the current populations; δ_s - relative time interval between t_s and the ancient split event (t_{As}); ϵ_{pool} - experimental error introduced by the pooling procedures; ϵ_{seq} - error associated with sequencing and mapping errors; m_{12}, m_{34} - probability per generation that an individual migrates from the N_1 or N_3 (Crab) population to the N_2 or N_4 (Wave) population (forward in time), m_{21}, m_{43} - probability per generation that an individual migrates from the N_2 or N_4 (Wave) population to the N_1 or N_3 (Crab) population (forward in time); $4N_2m_{12}$ and $4N_1m_{21}$ - average number of immigrants per generation ($4Nm$) from N_1 to N_2 and from N_2 to N_1 (respectively) at the first site; $4N_4m_{34}$ and $4N_3m_{43}$ - equivalent immigration rates at the second site and P_{no} - proportion of the simulated loci where no migration occurs between ecotypes.

parameter	prior mean	two-population	single origin	parallel origin
n_1	0.997	0.119	0.111	0.128
n_2	0.998	0.110	0.113	0.121
n_3	0.997	–	0.121	0.140
n_4	0.999	–	0.127	0.129
na_1	0.998	–	0.596	0.530
na_2	1.000	–	0.616	0.549
t_{div}	1.000	0.342	–	–
t_s	1.000	–	0.036	0.172
δ_s	1.001	–	0.182	0.179
ϵ_{pool}	1.000	0.242	0.243	0.241
ϵ_{seq}	1.001	0.592	0.062	0.042
m_{12}, m_{34}	1.000	0.401	0.396	0.448
m_{21}, m_{43}	1.001	0.448	0.399	0.439
$4N_2m_{12}$	0.999	0.284	0.325	0.311
$4N_1m_{21}$	0.998	0.293	0.287	0.329
$4N_4m_{34}$	0.996	–	0.298	0.319
$4N_3m_{43}$	1.000	–	0.298	0.340
P_{no}	1.000	0.072	0.041	0.124

Table 4: **Absolute parameter estimates for *Littorina saxatilis* populations.** Results are shown for the Arsklovet population for the two-population model and for Arsklovet and Ramsö for the single origin model. For this model N_1 and N_2 correspond, respectively, to the absolute size of the Arsklovet Crab and Wave populations, while N_3 and N_4 correspond to the absolute size of the Ramsö Crab and Wave populations, respectively. For each parameter, the value outside brackets corresponds to the re-scaled mean of the posterior distribution and in-between brackets is the 95% credible interval. T_{div}, T_s and Δ_s are presented in years. Parameters indicated here are the same as in table 3, except for P_{no}, which is converted to the percentage of the genome where no migration occurs between ecotypes.

parameter	two-population	single origin
N_1	18489 (12106 - 32956)	10336 (4617 - 34148)
N_2	15793 (10167 - 27613)	5486 (2936 - 18424)
N_3	–	12648 (5488 - 35603)
N_4	–	15309 (6245 - 41201)
NA_1	–	40854 (8516 - 53242)
NA_2	–	21118 (3866 - 47367)
T_{div}	18211 (2210 - 111264)	–
T_s	–	521 (316 - 818)
Δ_s	–	14308 (4790 - 42954)
$4N_2m_{12}$	22.8 (5.9 - 60.8)	30.6 (10.3 - 105.1)
$4N_1m_{21}$	16.3 (2.3 - 52.6)	32.1 (10.1 - 108.0)
$4N_4m_{34}$	–	34.3 (11.0 - 117.2)
$4N_3m_{43}$	–	19.9 (6.3 - 71.4)
P_{no}	1.2 (0.1 - 6.6)	1.3 (0.2 - 5.4)

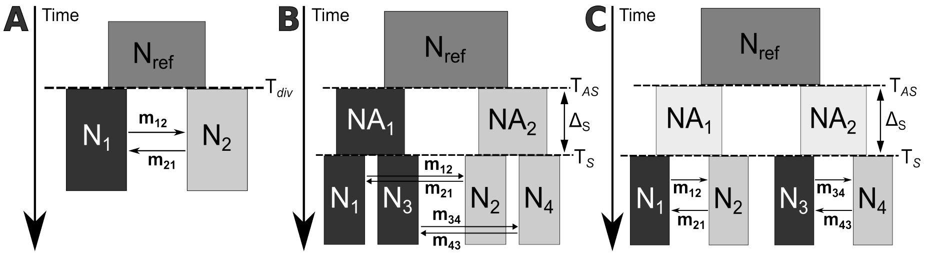

Figure 1: Demographic models for the isolation with migration scenario with two populations (A), single (B) and parallel (C) ecotype formation. Dark shading indicates one of the ecotypes, light shading the other ecotype. Parameters used were: N_{ref} - effective size of the ancestral population, NA_1 and NA_2 - size of the two ancestral populations, N_1 - N_4 - sizes of the present-day populations, T_{div} - time of separation of the ecotype populations (in generations), T_s - time of the recent split event (in generations), T_{As} - time of the ancient split event (in generations), Δ_s - time interval between the two split events (in generations), m_{12} - probability per generation that an individual migrates from N_1 to N_2 (forward in time), which corresponds to the probability that lineages move from N_2 to N_1 backwards in time, m_{21} - probability per generation that an individual migrates from N_2 to N_1 (forward in time), which corresponds to the probability that lineages move from N_1 to N_2 backwards in time, m_{34} - probability per generation that an individual migrates from N_3 to N_4 (forward in time), which corresponds to the probability that lineages move from N_4 to N_3 backwards in time and m_{43} - probability per generation that an individual migrates from N_4 to N_3 (forward in time), which corresponds to the probability that lineages move from N_3 to N_4 backwards in time.

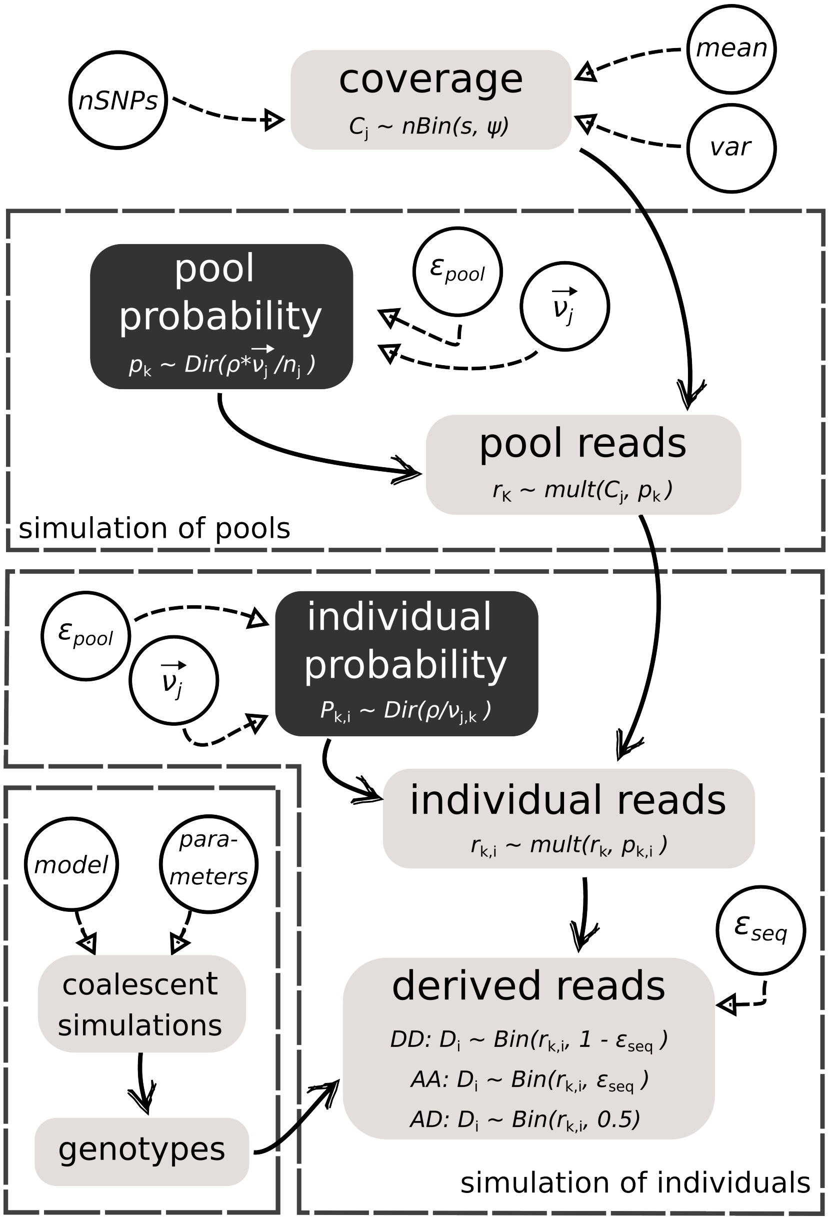
coverage
$C_j \sim nBin(s, \psi)$
nSNPs
mean
var
pool probability
$p_k \sim Dir(\rho * \vec{\nu}_j / n_j)$
ε_{pool}
$\vec{\nu}_j$
pool reads
$r_K \sim mult(C_j, p_k)$
simulation of pools
ε_{pool}
$\vec{\nu}_j$
individual probability
$P_{k,i} \sim Dir(\rho/\nu_{j,k})$
individual reads
$r_{k,i} \sim mult(r_k, p_{k,i})$
model
para-meters
coalescent simulations
genotypes
ε_{seq}
derived reads
DD: $D_i \sim Bin(r_{k,i}, 1 - \varepsilon_{seq})$
AA: $D_i \sim Bin(r_{k,i}, \varepsilon_{seq})$
AD: $D_i \sim Bin(r_{k,i}, 0.5)$
simulation of individuals

Figure 2: Schematics of the steps needed to simulated Pool-seq data. Dark colored boxes denote steps related with probabilities of contribution and circles represent necessary inputs for the corresponding step. Important formulas for each step are included inside the relevant box.

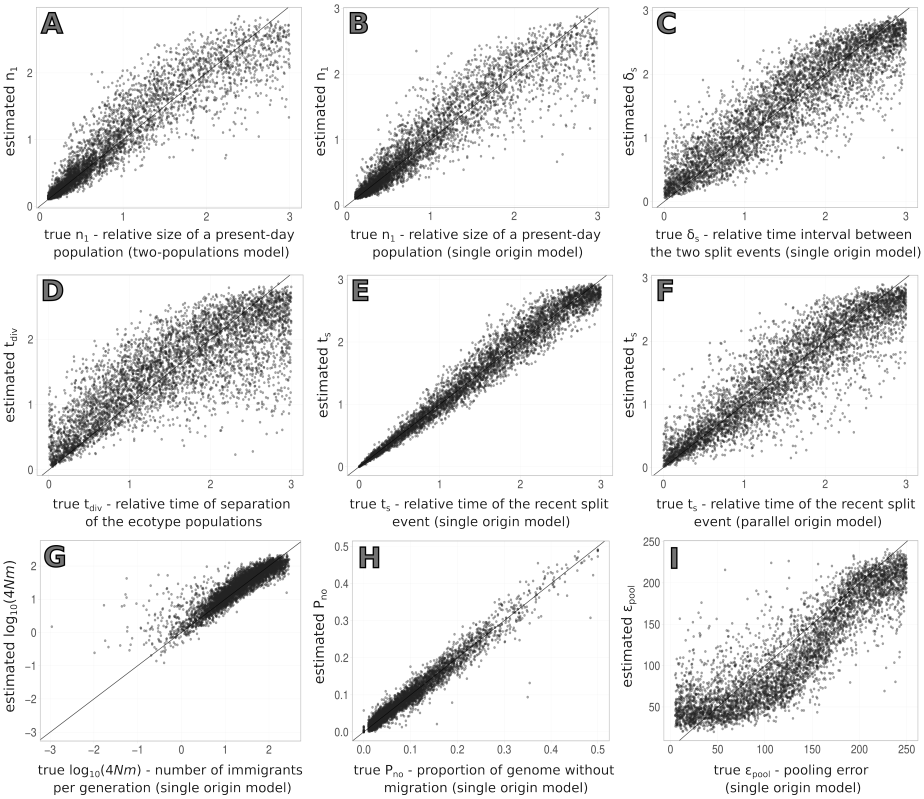

Figure 3: Results of the cross-validation for parameter estimation. The y-axis displays the estimated values, plotted against the true parameter values on the x-axis. Estimates correspond to the mean of the posterior obtained with a tolerance rate of 0.01. Parameters shown here are: A - relative size of a present-day population (n_1) of the two-population model; B - relative size of a present-day population (n_1) of the single origin model; C - time interval between the two split events (δ_s); D to F - time of the split event (t_{div}) for the two-population model and time of the recent split (t_s) for the single origin model and the parallel origin model (respectively); G - average number of immigrants per generation in log_{10} scale ($4Nm$); H - proportion of the genome without migration between different populations (P_{no}) and I - pooling error.

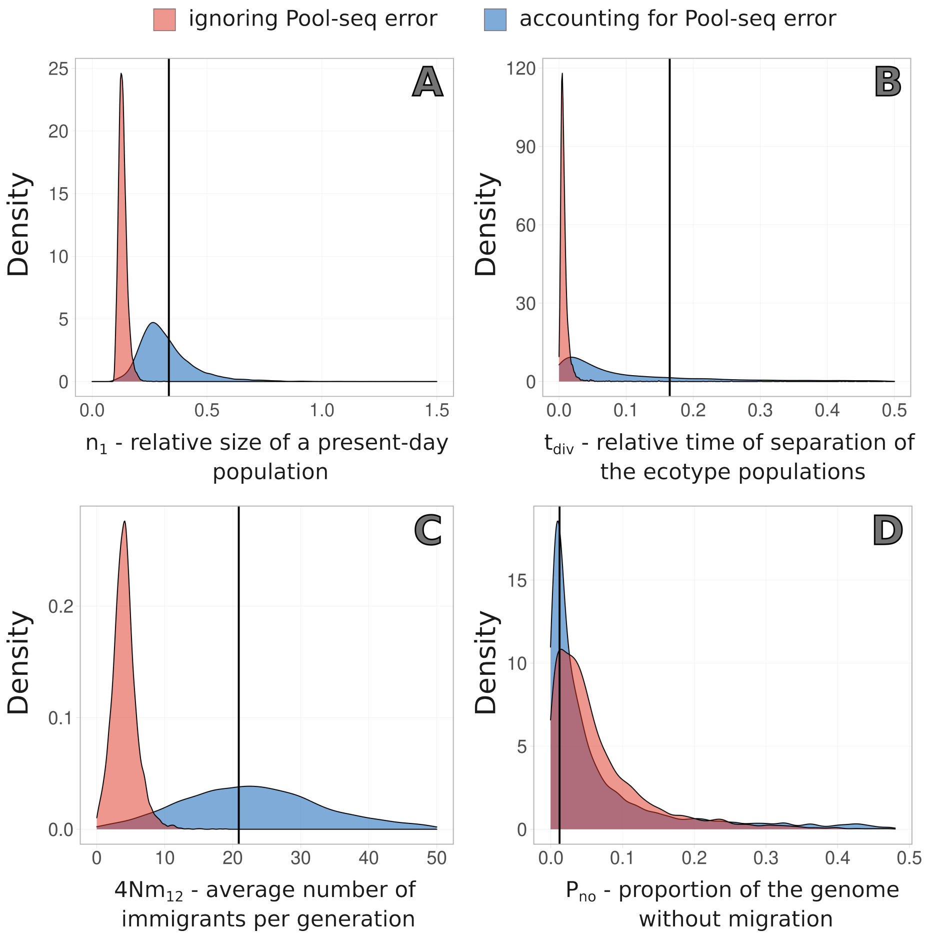

Figure 4: Impact of ignoring Pool-seq errors on demographic parameter estimates. Posterior obtained for a pseudo-observed Pool-seq dataset using either our ABC approach that explicitly accounts for Pool-seq errors (blue), or ignoring Pool-seq errors by using directly simulated allele frequencies (red). The parameters shown here are: A - relative size of a present-day population (n_1), B - relative time of separation of the ecotype populations (t_{div}), C - average number of immigrants per generation ($4Nm_{12}$) and D - proportion of the genome without migration (P_{no}). The black line represents the true parameter value used to simulate the pseudo-observed dataset with $L = 100$ loci.

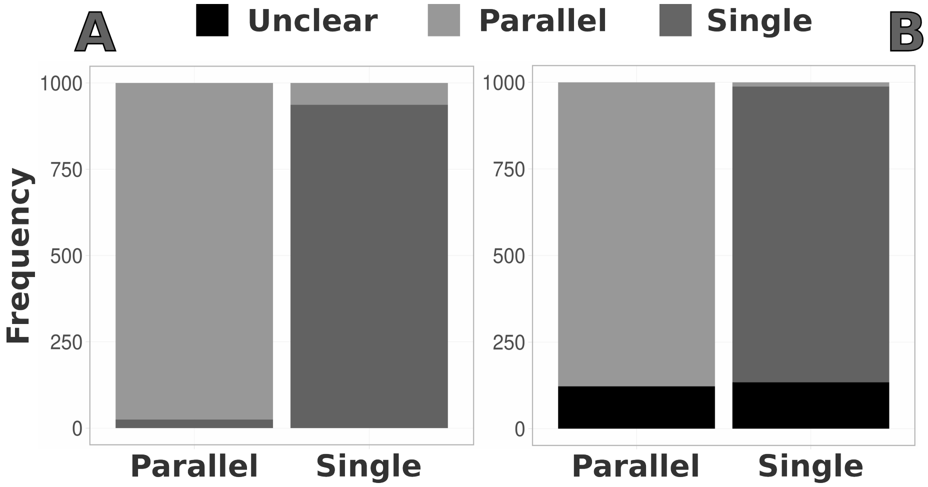

Figure 5: Model misclassification for the four-population models. Confusion matrix assuming that a simulation is assigned to a given model when the posterior probability is above 0.5 (A) or assuming that a simulation is only assigned to a model when the posterior probability is above 0.9 (B).

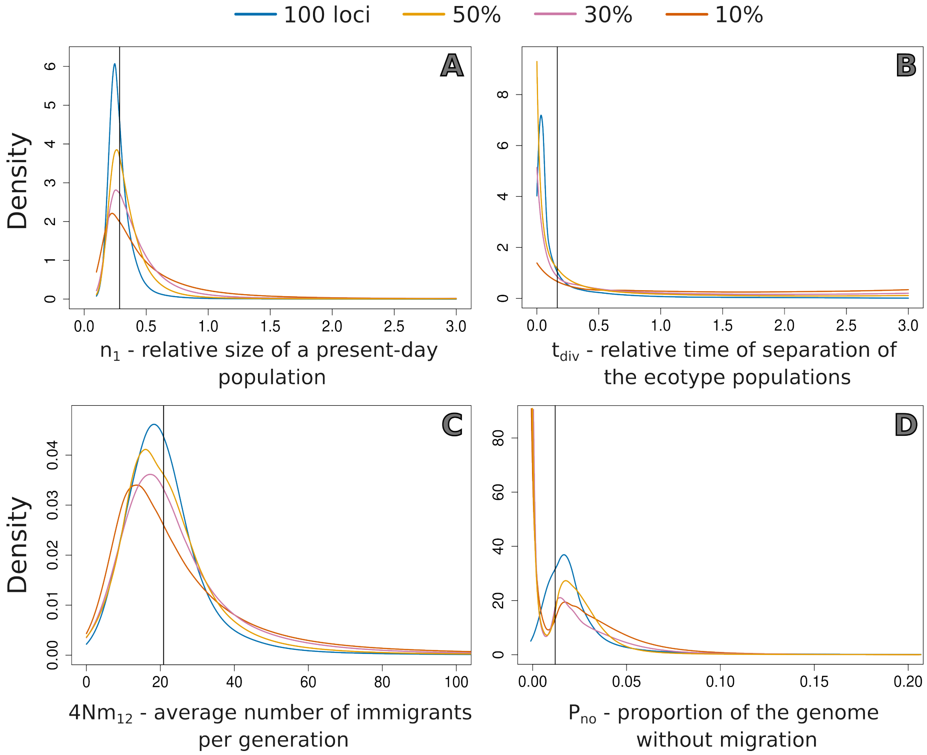

Figure 6: Impact of merging posteriors. We generated a pseudo-observed dataset of 100 loci and inferred parameters using the full dataset or subsets representing 10%, 30%, or 50% of the genome. The x-axis shows the estimated parameter value, and the y-axis shows the density of the posterior distribution obtained with the full dataset and the weighted combination of posteriors from the subsets. The solid vertical line represents the true parameter value. Parameters shown are: A - relative size of a present-day population (n_1), B - relative time of separation of the ecotype populations (t_{div}), C - average number of immigrants per generation ($4Nm_{12}$) and D - proportion of the genome without migration (P_{no}).

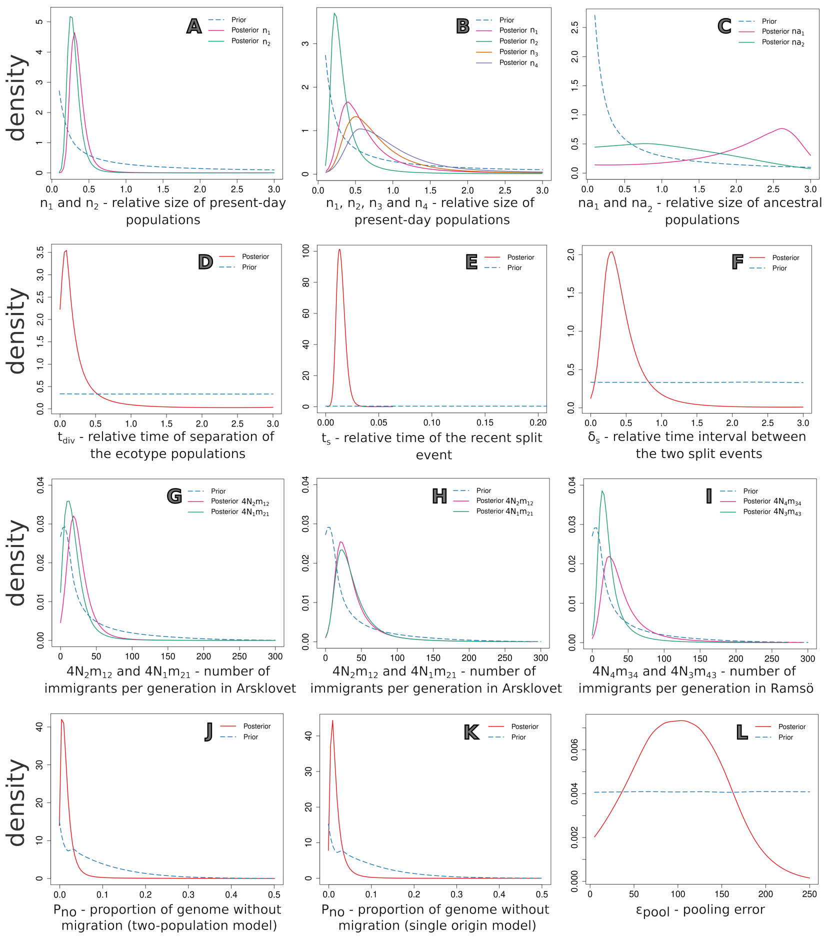

Figure 7: Posterior distributions of relative *L. saxatilis* parameters using regression adjustment method and a tolerance of 0.01. Prior distributions are shown for reference (dotted blue line). First column (A, D, G and J) corresponds to the two-population model, others to the single origin model. A - relative size of Arsklovet Crab (n_1) and Wave populations (n_2), B - relative size of Arsklovet Crab (n_1), Arsklovet Wave (n_2), Ramsö Crab (n_3) and Ramsö Wave (n_4) populations, C - relative size of ancestral populations (na_1 and na_2), D - relative time of separation of the ecotype populations (t_{div}), E - relative time of the recent split event (t_s), F - relative time interval between the two split events (δ_s), G and H - average number of immigrants per generation ($4N_2m_{CW}$ and $4N_1m_{WC}$) in Arsklovet, I - average number of immigrants per generation ($4N_4m_{CW}$ and $4N_3m_{WC}$) in Ramsö, J and K - proportion of the genome without migration (P_{no}) and L - pooling error. The relative parameter values were converted to absolute values using a re-scaling factor $f = obs[S]/E[S]$, where $obs[S]$ corresponds to the observed number of SNPs and $E[S]$ is the expected number of SNPs. Absolute parameter values were obtained by multiplying the point estimate of the posteriors shown here by the rescaling factor f.

Supplementary Material

Input and output files

For the simulation of Pool-seq data, our method relies on custom-made R functions that do not require a particular input file but instead require a set of user inputs at each appropriate function. To simulate the total depth of coverage for each population, the user must define the mean and the variance of the depth of coverage for each population, as well as the total number of SNPs to simulate. To simulate pools, the user must also define the pool error to use in the simulation (ϵ_{pool}). Finally, to obtain the number of reads with the derived allele, D_i, the user must also supply a value for the sequencing/mapping error (ϵ_{seq}) and the genotypes, ideally obtained using coalescent theory to simulate gene trees. After this step, our method provides a function to translate the number of ancestral/derived alleles into major/minor alleles, ensuring that the minor allele is the one for which we have fewer reads across all the populations. At this step, the user also has the choice to remove sites with fewer than x minor allele reads, where x is a user-defined threshold. The output of this section of our method are two different matrices, one containing the number of minor allele reads and the other containing the total depth of coverage. Both matrices are in the $nPop \times nSNP$ format, meaning that each row contains the information for a given population, while each column is a different site. These matrices can be used to compute allele frequencies and thus, calculate several summary statistics.

Our approximate Bayesian computation method is designed to work with the *_rc* files produced by the snp-frequency-diff.pl script from the PoPoolation2 suite (Kofler, Pandey, & Schlötterer, 2011). This file contains the number of major and minor allele reads for every SNP in a concise format (for more information please see the PoPoolation2 manual: `https://sourceforge.net/p/popoolation2/wiki/Manual/`). Given the modular nature of our method, it can also accommodate inputs in the form of matrices, where one of the matrices contains the number of minor allele reads and the other contains the total depth of coverage. These matrices should be in the format $nPop \times nSNP$, meaning that each row should contain the information for a given population, while each column is a different site. Note that an additional matrix, of the same dimensions, containing SNP position and contig information should also be available. The input files can then be filtered, removing sites with high or low coverage and sites with too few minor allele reads. The threshold for both the coverage filter and the number of minor allele reads are defined by the user. For ABC parameter inference and model selection, summary statistics are computed for several random blocks of windows (selected according to the contig information) and used as the target. The final output of model selection includes the proportion of accepted simulations for a model under a rejection algorithm and the posterior model probabilities of each model after a local linear regression adjustment. For parameter inference, the output includes the estimates under the rejection algorithm, the regression adjusted estimates if a local linear regression was performed and the median, mean, mode and 95% confidence interval of the weighted posteriors for each parameter.

Table S1: **Set of summary statistics considered.** The D-statistics combinations tested if there was more introgression between the divergent ecotypes at the same location or between the same ecotypes at different locations: for D-statistic 1, P1 was the Wave population in the first location (N_2), P2 was the Wave population in the second location (N_4) and P3 was the Crab population at the first location (N_1); for D-statistic 2, P1 was again the Wave population in the first location (N_2) but P2 was the Crab population in the second location (N_3) and P3 was the Crab population at the first location (N_1); for D-statistic 3, P1 was also the Wave population at the first location (N_2), P2 was the Crab population at the first location (N_1) and P3 was the Wave population at the second location (N_4). For all combinations, P4 was assumed to be an outgroup fixed, at all sites, for the major allele. Note that for the four-population models we only considered the proportion of SNPs with fixed differences between the two populations that inhabit the same location. For the proportion of exclusive SNPs, we also computed this per location i.e. checking if each site was segregating in one population but not in the other population inhabiting the same location, but we also computed the proportion of sites that were segregating in only one population and not in the other three.

summary statistic	two-population	four-populations
mean heterozygosity [1]	2 values (1 per population)	4 values (1 per population)
SD heterozygosity [1]	2 values (1 per population)	4 values (1 per population)
mean heterozygosity between populations [1]	1 pairwise value	6 pairwise values
SD heterozygosity between populations [1]	1 pairwise value	6 pairwise values
pairwise F_{ST} [2]	1 pairwise value	6 pairwise values
SD F_{ST} [2]	1 pairwise value	6 pairwise values
$5\% F_{ST}$ [2]	1 pairwise value	6 pairwise values
$95\% F_{ST}$ [2]	1 pairwise value	6 pairwise values
proportion of fixed differences [3]	1 pairwise value	2 values
proportion of exclusive SNPs [3]	2 values (1 per population)	5 values
mean D-statistic 1 [4]	–	1 value
mean D-statistic 2 [4]	–	1 value
mean D-statistic 3 [4]	–	1 value
SD D-statistic 1 [4]	–	1 value
SD D-statistic 2 [4]	–	1 value
SD D-statistic 3 [4]	–	1 value
total	13	57

[1] - Nei and Roychoudhury (1974); [2] - Bhatia et al. (2013); [3] - Fraïsse et al. (2021); [4] - Adapted from Malinsky et al. (2021) assuming that the outgroup was fixed for an allele different from P3, using $nABBA = \sum_{i=1}^{L}(p_{i1}(1-p_{i2})(1-p_{i3})) + ((1-p_{i1})p_{i2}p_{i3})$, $nBABA = \sum_{i=1}^{L}((1-p_{i1})p_{i2}(1-p_{i3})) + (p_{i1}(1-p_{i2})p_{i3})$, where pij denotes the minor-allele frequency at site i for population j.

Table S2: **Prediction errors for two-population model parameters.** Parameter inference was performed using a simple rejection or a regression adjustment using a local linear regression. For each method, values are presented for two different tolerance rates. n_1 and n_2 - relative population sizes of the extant populations, t_{div} - relative time of separation of the ecotype populations, ϵ_{pool} - experimental error introduced by the pooling procedures, ϵ_{seq} - error associated with sequencing and mapping errors, m_{12} - probability per generation that an individual migrates from N_1 to N_2 (forward in time), m_{21} - probability per generation that an individual migrates from N_2 to N_1 (forward in time), $4N_2m_{12}$ and $4N_1m_{21}$ - average number of immigrants per generation ($4Nm$) from N_1 to N_2 and from N_2 to N_1 (respectively) and P_{no} - proportion of the simulated loci where no migration occurs between ecotypes.

	REJECTION						**REGRESSION**					
	tolerance of 0.005			*tolerance of 0.01*			*tolerance of 0.005*			*tolerance of 0.01*		
parameter	mode	median	mean	mode	median	mean	mode	median	mean	mode	median	mean
n_1	0.312	0.219	0.220	0.349	0.243	0.243	0.113	0.106	0.106	0.127	0.119	0.119
n_2	0.310	0.213	0.213	0.356	0.239	0.239	0.118	0.110	0.110	0.120	0.111	0.110
t_{div}	1.023	0.564	0.589	1.225	0.607	0.634	0.456	0.340	0.319	0.500	0.367	0.342
ϵ_{pool}	0.625	0.414	0.432	0.658	0.461	0.487	0.261	0.239	0.236	0.262	0.242	0.242
ϵ_{seq}	2.527	0.966	0.974	2.651	0.974	0.981	0.914	0.613	0.591	0.884	0.611	0.592
m_{12}	1.404	0.648	0.674	1.390	0.651	0.687	0.655	0.436	0.428	0.609	0.402	0.401
m_{21}	1.301	0.627	0.656	1.368	0.668	0.697	0.668	0.432	0.423	0.710	0.462	0.448
$4N_2m_{12}$	0.762	0.501	0.485	0.800	0.528	0.512	0.338	0.287	0.283	0.336	0.286	0.284
$4N_1m_{21}$	0.768	0.486	0.466	0.837	0.555	0.525	0.308	0.262	0.259	0.351	0.297	0.293
P_{no}	0.323	0.233	0.214	0.327	0.232	0.212	0.117	0.102	0.090	0.089	0.080	0.072

Table S3: **Prediction errors for the single origin parameters.** Parameter inference was performed using a simple rejection or a regression adjustment using a local linear regression. For each method, values are presented for two different tolerance rates. n_1 to n_4 - relative population sizes of the extant populations, na_1 and na_2 - relative population sizes of the ancestral populations, t_s - relative time of the split event that lead to the origin of the current populations, δ_s - relative time interval between t_s and the ancient split event (t_{As}), ϵ_{pool} - experimental error introduced by the pooling procedures, ϵ_{seq} - error associated with sequencing and mapping errors, m_{12}, m_{34} - probability per generation that an individual migrates from the N_1 or N_3 (Crab) population to the N_2 or N_4 (Wave) population (forward in time), m_{21}, m_{43} - probability per generation that an individual migrates from the N_2 or N_4 (Wave) population to the N_1 or N_3 (Crab) population (forward in time), $4N_2m_{12}$ and $4N_1m_{21}$ - average number of immigrants per generation ($4Nm$) from N_1 to N_2 and from N_2 to N_1 (respectively) at the first site, $4N_4m_{34}$ and $4N_3m_{43}$ - equivalent immigration rates at the second site and P_{no} - proportion of the simulated loci where no migration occurs between ecotypes.

	REJECTION						**REGRESSION**					
	tolerance of 0.005			*tolerance of 0.01*			*tolerance of 0.005*			*tolerance of 0.01*		
parameter	mode	median	mean	mode	median	mean	mode	median	mean	mode	median	mean
n_1	0.759	0.465	0.417	0.830	0.489	0.447	0.142	0.127	0.122	0.126	0.114	0.111
n_2	0.857	0.513	0.451	0.934	0.546	0.490	0.138	0.123	0.119	0.133	0.118	0.113
n_3	0.734	0.452	0.409	0.880	0.530	0.474	0.126	0.113	0.110	0.140	0.125	0.121
n_4	0.821	0.501	0.448	0.957	0.563	0.495	0.127	0.115	0.112	0.149	0.134	0.127
na_1	1.949	1.109	0.954	1.945	1.119	0.963	1.316	0.613	0.583	1.407	0.627	0.596
na_2	1.943	1.103	0.955	1.933	1.112	0.963	1.383	0.643	0.615	1.415	0.646	0.616
t_s	0.070	0.063	0.067	0.075	0.071	0.078	0.039	0.037	0.036	0.039	0.037	0.036
δ_s	1.327	0.694	0.734	1.452	0.741	0.778	0.228	0.193	0.185	0.223	0.188	0.182
ϵ_{pool}	1.256	0.704	0.767	1.429	0.766	0.822	0.266	0.253	0.236	0.271	0.261	0.243
ϵ_{seq}	0.539	0.550	0.629	0.619	0.627	0.703	0.084	0.070	0.062	0.088	0.071	0.062
m_{12}, m_{34}	1.579	0.744	0.794	1.569	0.781	0.827	0.523	0.386	0.379	0.559	0.401	0.396
m_{21}, m_{43}	1.410	0.738	0.790	1.528	0.798	0.842	0.522	0.384	0.377	0.549	0.401	0.399
$4N_2m_{12}$	1.072	0.759	0.659	1.156	0.843	0.720	0.357	0.299	0.276	0.426	0.357	0.325
$4N_1m_{21}$	1.113	0.773	0.657	1.123	0.808	0.709	0.396	0.330	0.299	0.367	0.307	0.287
$4N_4m_{34}$	1.129	0.811	0.696	1.188	0.865	0.731	0.365	0.306	0.280	0.393	0.328	0.298
$4N_3m_{43}$	1.153	0.818	0.687	1.149	0.840	0.727	0.358	0.299	0.274	0.388	0.323	0.298
P_{no}	0.190	0.135	0.125	0.235	0.162	0.149	0.044	0.042	0.041	0.045	0.043	0.041

Table S4: **Prediction errors for the parallel origin parameters.** Parameter inference was performed using a simple rejection or a regression adjustment using a local linear regression. For each method, values are presented for two different tolerance rates. n_1 to n_4 - relative population sizes of the extant populations, na_1 and na_2 - relative population sizes of the ancestral populations, t_s - relative time of the split event that lead to the origin of the current populations, δ_s - relative time interval between t_s and the ancient split event (t_{As}), ϵ_{pool} - experimental error introduced by the pooling procedures, ϵ_{seq} - error associated with sequencing and mapping errors, m_{12}, m_{34} - probability per generation that an individual migrates from the N_1 or N_3 (Crab) population to the N_2 or N_4 (Wave) population (forward in time), m_{21}, m_{43} - probability per generation that an individual migrates from the N_2 or N_4 (Wave) population to the N_1 or N_3 (Crab) population (forward in time), $4N_2m_{12}$ and $4N_1m_{21}$ - average number of immigrants per generation ($4Nm$) from N_1 to N_2 and from N_2 to N_1 (respectively) at the first site, $4N_4m_{34}$ and $4N_3m_{43}$ - equivalent immigration rates at the second site and P_{no} - proportion of the simulated loci where no migration occurs between ecotypes.

	REJECTION						**REGRESSION**					
	tolerance of 0.005			*tolerance of 0.01*			*tolerance of 0.005*			*tolerance of 0.01*		
parameter	mode	median	mean	mode	median	mean	mode	median	mean	mode	median	mean
n_1	0.743	0.437	0.395	0.888	0.512	0.455	0.157	0.138	0.131	0.149	0.134	0.128
n_2	0.731	0.444	0.397	0.858	0.497	0.445	0.145	0.128	0.123	0.141	0.126	0.121
n_3	0.910	0.525	0.456	0.968	0.563	0.494	0.154	0.136	0.130	0.171	0.149	0.140
n_4	0.836	0.480	0.423	0.961	0.553	0.487	0.156	0.137	0.129	0.153	0.135	0.129
na_1	1.913	0.899	0.806	1.914	0.942	0.834	1.161	0.562	0.533	1.199	0.560	0.530
na_2	1.925	0.923	0.813	1.958	0.962	0.840	1.116	0.547	0.524	1.194	0.582	0.549
t_s	0.549	0.360	0.385	0.603	0.389	0.415	0.204	0.171	0.158	0.223	0.189	0.172
δ_s	0.474	0.334	0.347	0.493	0.353	0.367	0.202	0.176	0.167	0.212	0.188	0.179
ϵ_{pool}	1.270	0.710	0.760	1.346	0.753	0.801	0.263	0.245	0.240	0.261	0.244	0.241
ϵ_{seq}	0.531	0.471	0.539	0.600	0.542	0.611	0.061	0.051	0.044	0.061	0.049	0.042
m_{12}, m_{34}	1.505	0.767	0.809	1.582	0.803	0.843	0.609	0.454	0.447	0.606	0.447	0.448
m_{21}, m_{43}	1.514	0.781	0.822	1.570	0.800	0.841	0.606	0.450	0.443	0.580	0.438	0.439
$4N_2m_{12}$	1.118	0.772	0.658	1.161	0.799	0.685	0.387	0.331	0.311	0.396	0.333	0.311
$4N_1m_{21}$	1.119	0.764	0.658	1.184	0.827	0.704	0.407	0.345	0.319	0.417	0.353	0.329
$4N_4m_{34}$	1.150	0.789	0.653	1.201	0.860	0.726	0.417	0.345	0.331	0.400	0.340	0.319
$4N_3m_{43}$	1.179	0.834	0.667	1.208	0.870	0.733	0.412	0.350	0.326	0.432	0.367	0.340
P_{no}	0.484	0.314	0.278	0.579	0.368	0.327	0.129	0.120	0.118	0.134	0.126	0.124

Table S5: **Biases of the estimates obtained when explicitly modeling or ignoring Pool-seq errors.** We simulated pseudo-observed Pool-seq data and inferred parameters using either a table of summary statistics computed directly from simulated haplotypes without accounting for Pool-seq errors or a table of summary statistics computed after simulating Pool-seq data and explicitly considering depth of coverage variation, unequal individual contribution, and sequencing errors. We computed the bias of the estimates using $\frac{1}{n} \cdot \sum(|\hat{\Theta}_i - \Theta_i|)$, where $\hat{\Theta}_i$ is the estimated mean posterior, and Θ_i is the true parameter value for the i^{th} pseudo-observed dataset, while $n = 100$ is the number of simulated pseudo-observed datasets. n_1 and n_2 - relative population sizes of the present-day populations, t_{div} - relative time of separation of the ecotype populations, $4Nm_{12}$ and $4Nm_{21}$ - average number of immigrants per generation and P_{no} - proportion of the genome without migration.

parameter	ignoring Pool-seq data	accounting for Pool-seq data
n_1	0.605	0.144
n_2	0.584	0.192
t_{div}	0.939	0.630
$4Nm_{12}$	0.797	0.187
$4Nm_{21}$	0.719	0.239
P_{no}	0.018	0.010

Table S6: **Biases of the estimates obtained with subsets of loci.** We simulated a pseudo-observed dataset of 100 loci and inferred parameters using the full dataset or subsets representing 10%, 30%, or 50% of the genome. To compute the bias, we contrasted the mean of the posterior distribution obtained with subsets of loci with the mean posterior obtained with 100 loci. The bias was computed a) after weighted combination of posteriors obtained with subsets representing 10%, 30% or 50% of the genome and b) by using the summary statistics of the full dataset as the target for parameter inference performed with 10%, 30% or 50% of the genome. n_1 and n_2 - relative population sizes of the present-day populations, t_{div} - relative time of separation of the ecotype populations, $4Nm_{12}$ and $4Nm_{21}$ - average number of immigrants per generation and P_{no} - proportion of the genome without migration.

	a) merging posteriors			b) whole-genome		
parameter	10%	30%	50%	10%	30%	50%
n_1	0.136	0.043	0.008	0.256	0.043	0.006
n_2	0.186	0.093	0.046	0.209	0.080	0.042
t_{div}	0.867	0.521	0.242	0.845	0.426	0.212
$4Nm_{12}$	4.564	1.584	1.863	13.261	0.730	1.878
$4Nm_{21}$	5.061	1.479	0.107	25.674	1.052	0.106
P_{no}	0.009	0.012	0.010	-0.020	0.003	0.006

Table S7: **Estimates for relative parameters of *Littorina saxatilis* populations.** Results are shown for the Arsklovet population for the two-population model and for Arsklovet and Ramsö for the single origin and parallel origin models. For these models n_1 and n_2 correspond to the Arsklovet Crab and Wave population respectively, while n_3 and n_4 correspond to the Ramsö Crab and Wave population respectively. For each parameter, the value outside brackets corresponds to the mean of the posterior distribution and in-between brackets is the 95% credible interval. Parameters here are the same as in table 2.

parameter	two-population	single origin	parallel origin
n_1	0.334 (0.219 - 0.596)	0.557 (0.249 - 1.841)	0.315 (0.134 - 0.732)
n_2	0.286 (0.184 - 0.499)	0.296 (0.158 - 0.993)	0.754 (0.241 - 1.895)
n_3	–	0.682 (0.296 - 1.919)	0.662 (0.208 - 1.718)
n_4	–	0.825 (0.337 - 2.221)	0.939 (0.277 - 2.189)
na_1	–	2.203 (0.459 - 2.870)	2.641 (1.554 - 2.980)
na_2	–	1.139 (0.208 - 2.554)	2.396 (0.873 - 2.963)
t_{div}	0.165 (0.020 - 1.517)	–	–
t_s	–	0.014 (0.009 - 0.022)	0.007 (0.005 - 0.018)
δ_s	–	0.386 (0.129 - 1.158)	0.029 (0.002 - 0.070)
m_{12}, m_{34}	0.00073 (0.00013 - 0.0009)	0.00048 (0.00012 - 0.00094)	0.00024 (0.00002 - 0.00076)
m_{21}, m_{43}	0.00049 (0.00005 - 0.00096)	0.00058 (0.00016 - 0.00096)	0.00077 (0.00028 - 0.00099)
P_{no}	0.012 (0.001 - 0.066)	0.013 (0.002 - 0.054)	0.205 (0.015 - 0.428)
ϵ_{pool}	182 (67 - 236)	102 (24 - 183)	130 (23 - 222)
ϵ_{seq}	0.00100 (0.00098 - 0.00100)	0.00092 (0.00059 - 0.00099)	0.00099 (0.00097 - 0.00100)

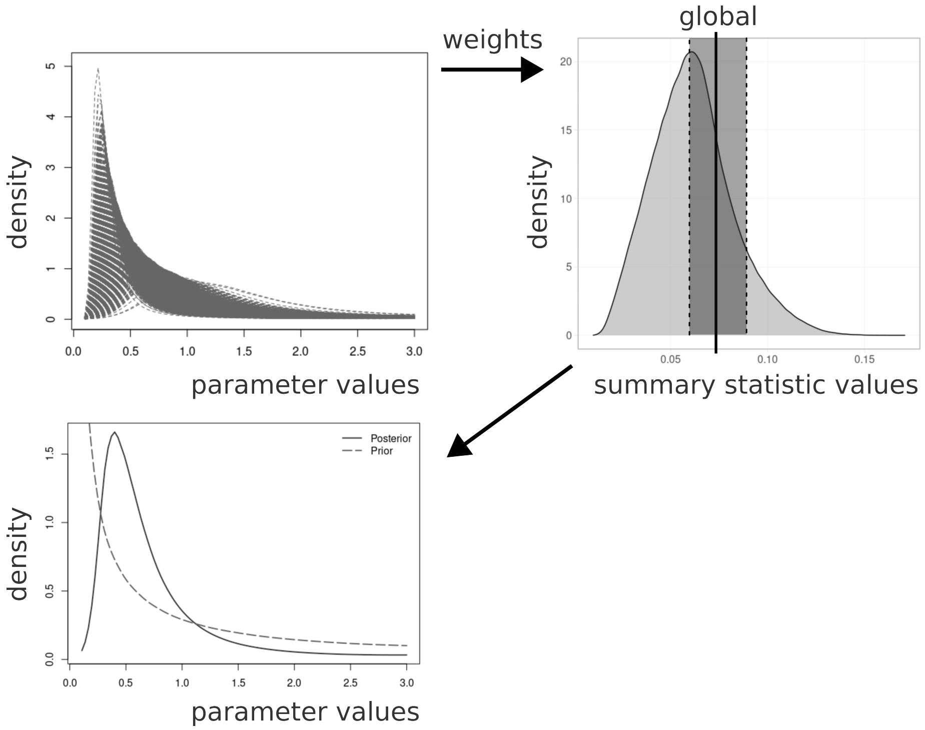

Figure S1: Merging of multiple posterior distributions. This represents an example of how the posteriors obtained for each set of loci are combined to obtain a single estimate per parameter. In the top-left plot several posteriors distributions are shown, one for each set of loci and for a given parameter. These multiple posteriors are weighted according to the distance between the summary statistics of the corresponding simulations and the mean across the genome, giving more weight to sets of loci with a mean closer to the overall mean. The top-right plot represents an example of this, where the simulations with values closer to the global value (represented by the black line) will have more weight. Using these weights, the multiple posteriors are combined to obtain a single estimate per parameter, as shown in the bottom-left panel.

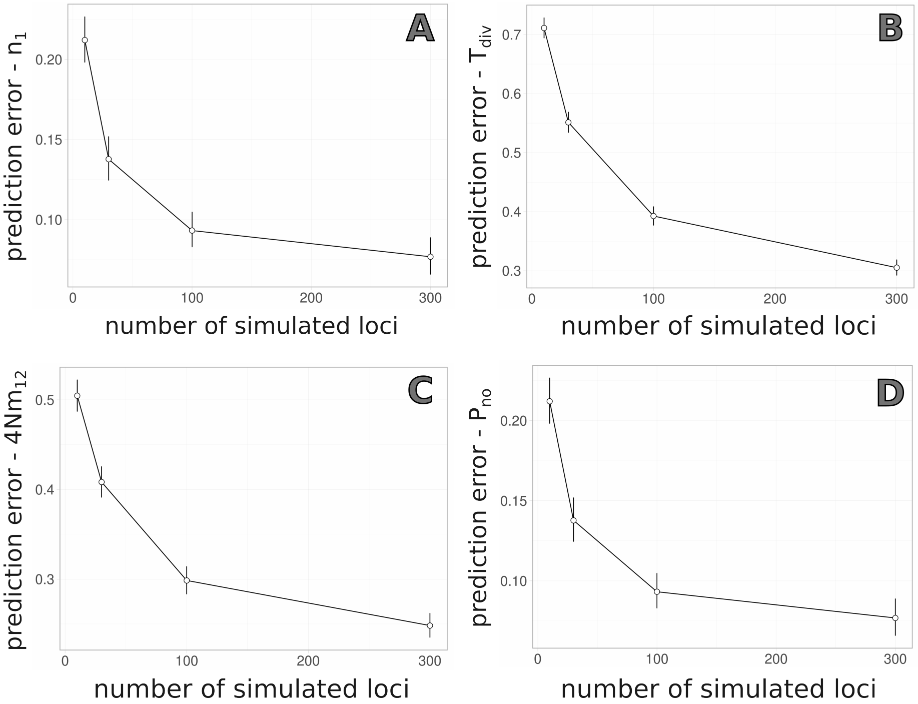

Figure S2: Impact of number of loci on the prediction error. A leave-one-out cross-validation simulation study with varying numbers of loci per subset was performed to compute the prediction error for several demographic parameters. The prediction error is shown on the y-axis. The x-axis shows the numbers of loci per subset. Points represent the mean prediction error after bootstrapping and error bars represent 95% confidence intervals. Parameters shown here are: A - relative size of a present-day population (n_1), B - relative time of separation of the ecotype populations (t_{div}), C - average number of immigrants per generation ($4Nm_{12}$) and D - proportion of the genome without migration (P_{no}).

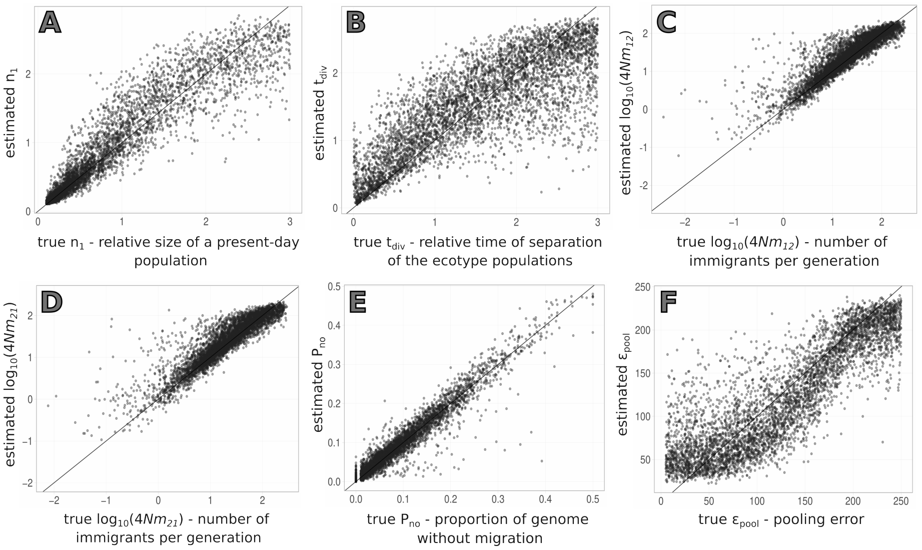

Figure S3: Results of the cross-validation for parameter estimation using the two-population model. The y-axis displays the estimated values, plotted against the true parameter values on the x-axis. Estimates correspond to the mean of the posterior obtained with a tolerance rate of 0.01. Parameters shown here are: A - relative size of a present-day population (n_1), B - relative time of separation of the ecotype populations (t_{div}), C and D - average number of immigrants per generation ($4Nm_{CW}$ and $4Nm_{WC}$, respectively), E - proportion of the genome without migration between different populations (P_{no}) and F - pooling error

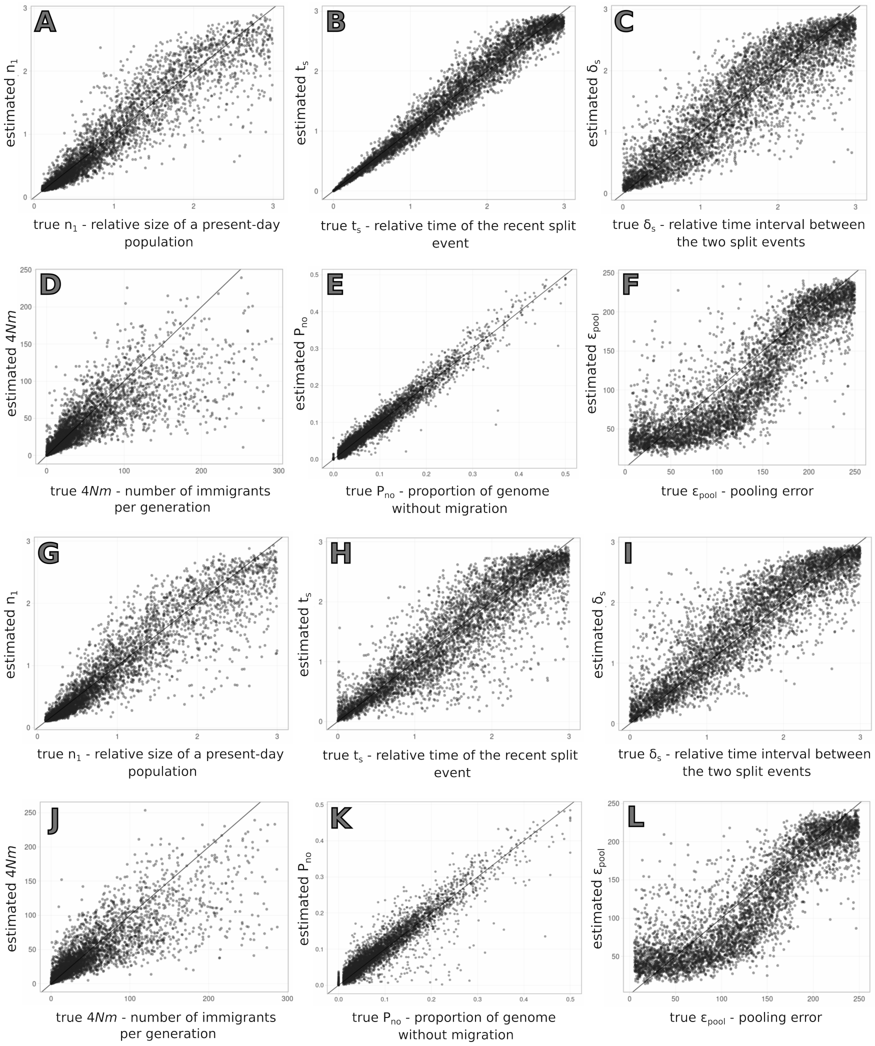

Figure S4: Results of the cross-validation for parameter estimation using the four-population models. Panels from A to F show the results for the single origin model, while panels G to L show the results for the parallel origin model. The y-axis displays the estimated values, plotted against the true parameter values on the x-axis. Estimates correspond to the mean of the posterior obtained with a tolerance rate of 0.01. Parameters shown here are: A and G - relative size of a present-day population (n_1), B and H - relative time of the recent split event (t_s), C and I - relative time interval between the two split events (δ_s), D and J - average number of immigrants per generation ($4Nm$), E and K - proportion of the genome without migration between different populations (P_{no}) and F and L - pooling error

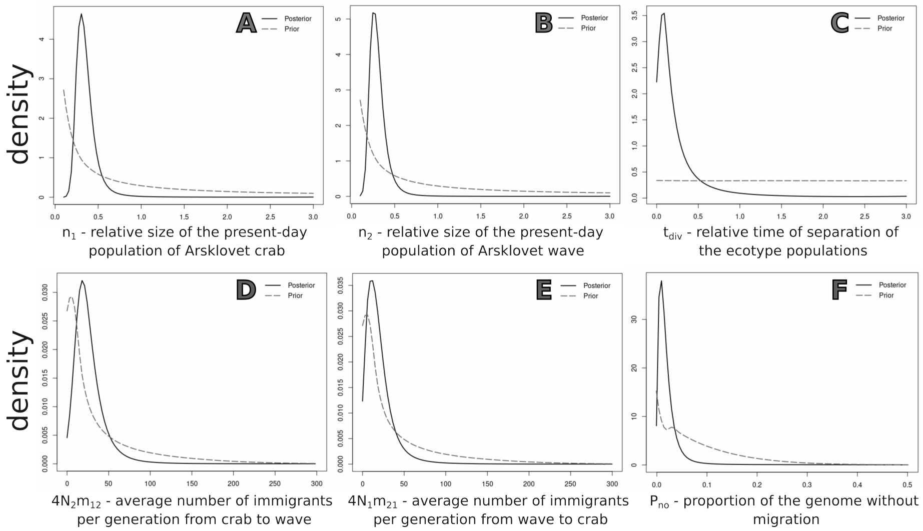

Figure S5: Posterior distribution of relative *L. saxatilis* parameters using the two-population model. The posterior distributions were obtained with the regression adjustment method and using a tolerance rate of 0.01. For reference, the prior distribution of each parameter is shown (dotted line). Parameters shown here are: A - relative size of the Arsklovet Crab population (n_1), B - relative size of the Arsklovet Wave population (n_2), C - relative time of separation of the ecotype populations (t_{div}), D and E - average number of immigrants per generation ($4Nm_{CW}$ and $4Nm_{WC}$, respectively) and F - proportion of the genome without migration between different populations (P_{no}). The relative parameter values presented here were converted to absolute values using a re-scaling factor $f = obs[S]/E[S]$, where $obs[S]$ corresponds to the observed number of SNPs and $E[S]$ is the expected number of SNPs. Absolute parameter values were obtained by multiplying the point estimate of the posteriors shown here by the rescaling factor f.

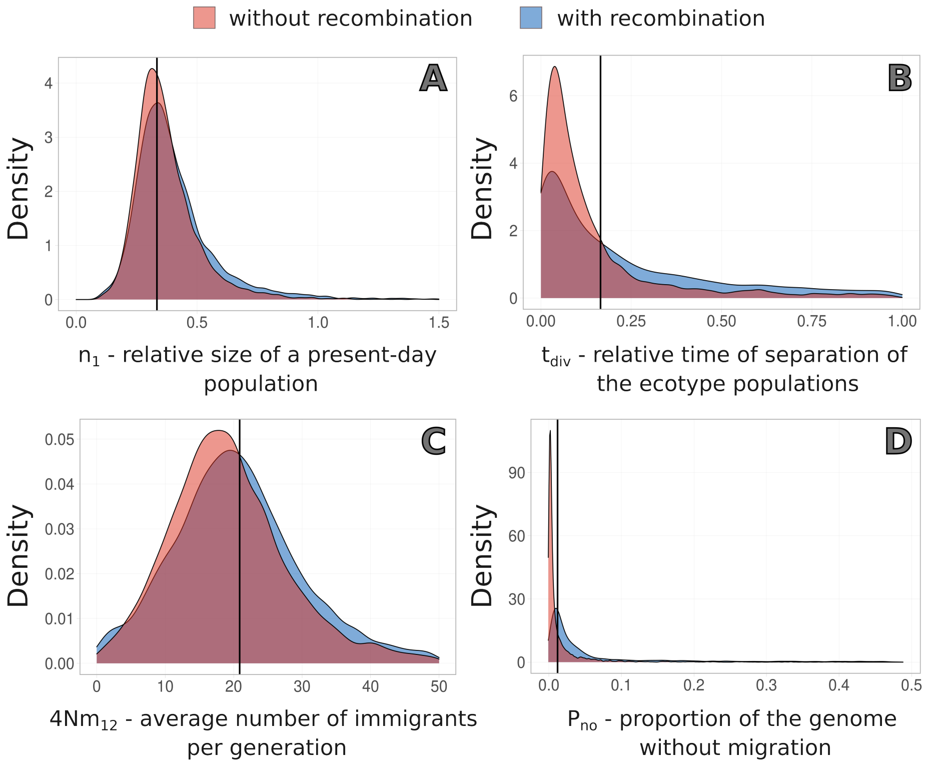

Figure S6: Impact of within-locus recombination on parameter estimates. We used simulations that excluded within-locus recombination to estimate the parameters of two pseudo-observed datasets: one with and another without within-locus recombination. The x-axis shows the estimated parameter value, and the y-axis shows the density of the posterior distribution. The posterior obtained for the pseudo-observed dataset without within-locus recombination is shown in red and the posterior for the pseudo-observed dataset with within-locus recombination in blue. The solid vertical line represents the true parameter value. Parameters shown here are: A - relative size of a present-day population (n_1), B - relative time of separation of the ecotype populations (t_{div}), C - average number of immigrants per generation ($4Nm_{12}$) and D - proportion of the genome without migration (P_{no}).

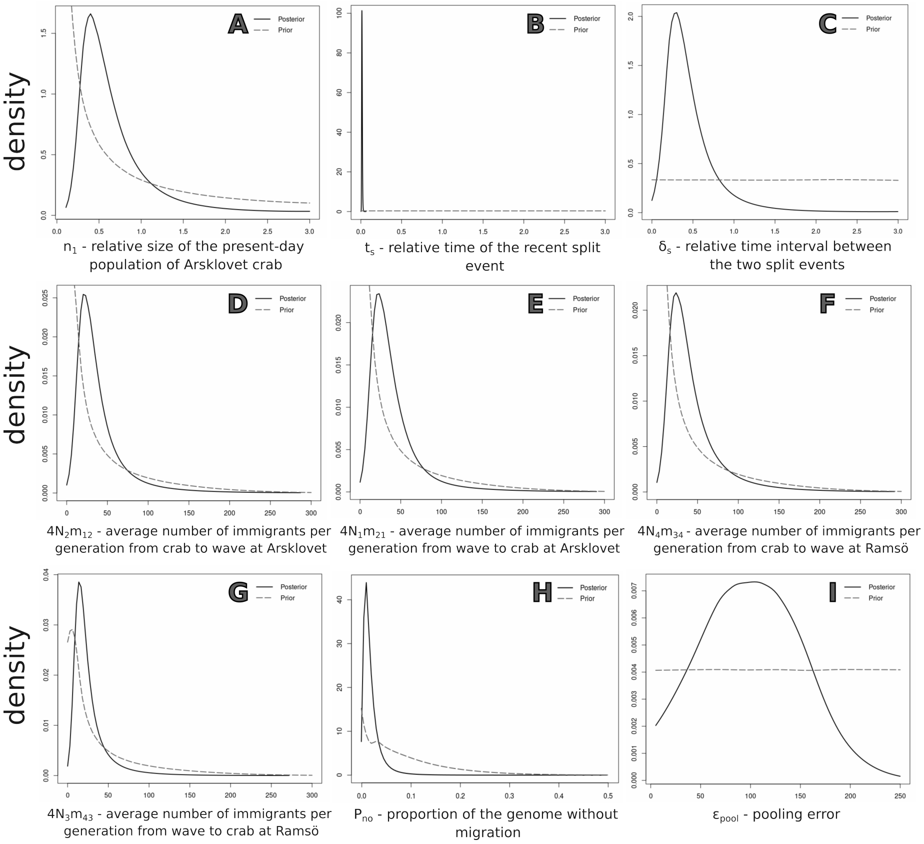

Figure S7: Posterior distribution of relative *L. saxatilis* parameters using the single origin model. The posterior distributions were obtained with the regression adjustment method and using a tolerance rate of 0.01. For reference, the prior distribution of each parameter is shown (dotted line). Parameters shown here are: A - relative size of the Arsklovet Crab population (n_1), B - relative time of the recent split event (t_s), C - relative time interval between the two split events (δ_s), D and E - average number of immigrants per generation ($4Nm$) from Crab to Wave and from Wave to Crab (respectively) at Arsklovet, F and G - average number of immigrants per generation ($4Nm$) from Crab to Wave and from Wave to Crab (respectively) at Ramsö, H - proportion of the genome without migration between different populations (P_{no}) and I - pooling error. The relative parameter values presented here were converted to absolute values using a re-scaling factor $f = obs[S]/E[S]$, where $obs[S]$ corresponds to the observed number of SNPs and $E[S]$ is the expected number of SNPs. Absolute parameter values were obtained by multiplying the point estimate of the posteriors shown here by the rescaling factor f.

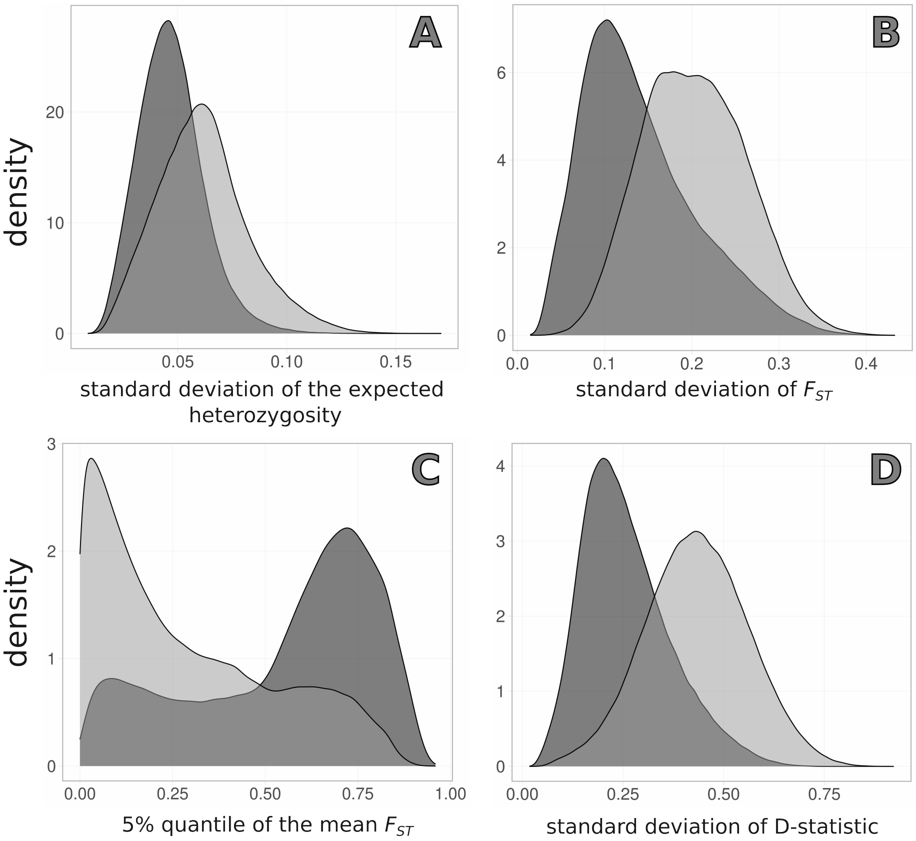

Figure S8: Distribution of summary statistics obtained for the single and parallel origin models. Dark shading indicates the parallel origin model and light shading the single origin model. Summary statistics are: A - standard deviation of the expected heterozygosity for a given population, B - standard deviation of mean pairwise F_{ST}, C - 5% quantile of the mean pairwise F_{ST} and D - standard deviation of D-statistic

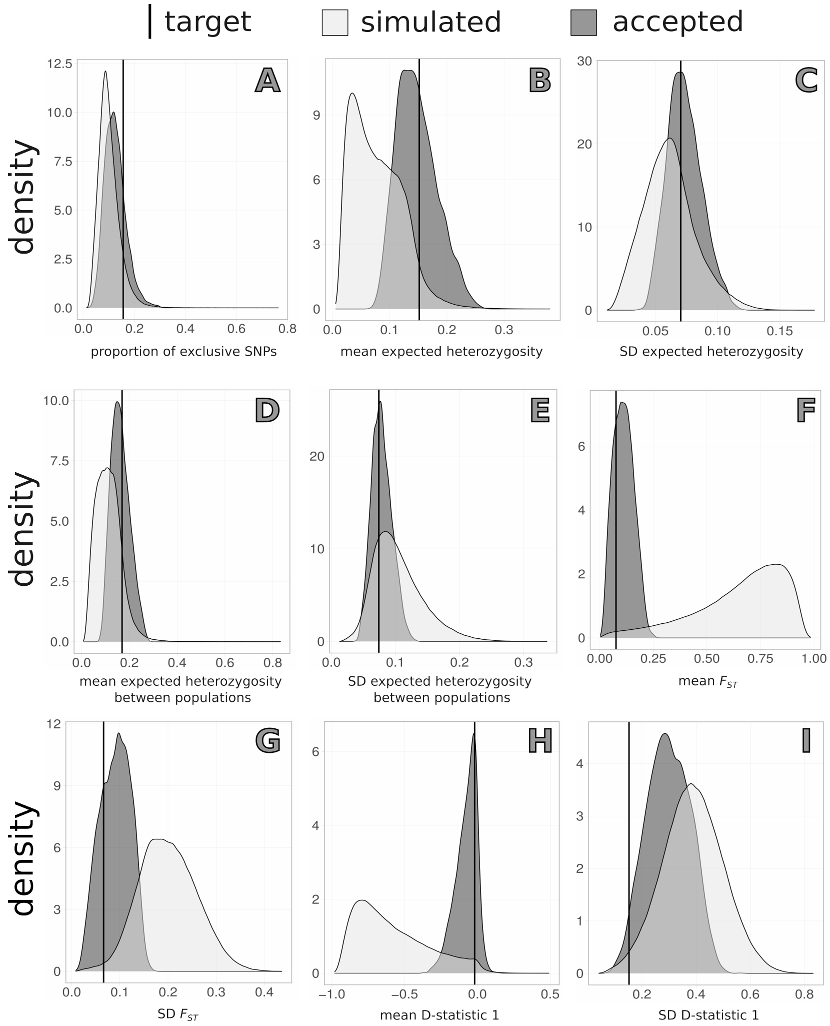

Figure S9: Distribution of accepted summary statistics. The black line represents the target for the parameter inference, the light shading is the distribution of the complete set of simulated summary statistics and the dark shading is the distribution of the accepted summary statistics for that particular target. Summary statistics include examples of all those analyzed here: A - proportion of exclusive sites, B - mean heterozygosity, C - standard deviation of the mean heterozygosity, D - mean heterozygosity between a pair of populations, E - standard deviation of the mean heterozygosity between a pair of populations, F - mean F_{ST} between a pair of populations, G - standard deviation of F_{ST}, H - D-statistic and I - standard deviation of D-statistic.