Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data

Thomas Guillemaud*, Mark A. Beaumont[†], Marc Ciosi[*], Jean-Marie Cornuet[‡§] and Arnaud Estoup[‡]

[*] Equipe "Biologie des Populations en Interaction", UMR 1301 IBSV INRA-CNRS-Université de Nice-Sophia Antipolis, Sophia Antipolis, France

[†] School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading, UK

[‡] INRA UMR Centre de Biologie et de Gestion des Populations (INRA / IRD / Cirad / Montpellier SupAgro), Campus international de Baillarguet, Montferrier-sur-Lez, France

[§] Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, London, U.K.

Running title: Introduction routes of invasive species

Keywords: ABC, model selection, bioinvasion, invasive species

Corresponding author: Thomas Guillemaud,

Equipe "Biologie des Populations en Interaction", UMR 1301 Interactions Biotiques et Santé Végétale, INRA-CNRS-Université de Nice-Sophia Antipolis.

400 Route des Chappes

BP 167 - 06903 Sophia Antipolis cedex FRANCE

E-mail: guillem@sophia.inra.fr

phone # 33 4 92 38 64 81, fax # 33 4 92 38 64 01

Word count in the text: 6869

Abstract

Determining the routes of introduction provides not only information about the history of an invasion process, but also information about the origin and construction of the genetic composition of the invading population. It remains difficult, however, to infer introduction routes from molecular data, due to a lack of appropriate methods. We evaluate here the use of an approximate Bayesian computation (ABC) method for estimating the probabilities of introduction routes of invasive populations based on microsatellite data. We considered the crucial case of a single source population from which two invasive populations originated either serially from a single introduction event or from two independent introduction events. Using simulated datasets, we found that the method gave correct inferences and was robust to many erroneous beliefs. The method was also more efficient than traditional methods based on raw values of statistics such as assignment likelihood or pairwise F_{ST}. We illustrate some of the features of our ABC method, using real microsatellite datasets obtained for invasive populations of the western corn rootworm, *Diabrotica virgifera virgifera*. Most computations were performed with the DIYABC program (http://www.montpellier.inra.fr/CBGP/diyabc/).

INTRODUCTION

In biological invasions, a large proportion of the genetic variability of an invading population depends on the historical and demographic features of its introduction: the number and genetic composition of sources, the number of successive introduction events from each source, the number of introduced individuals, the number of intermediate introduction steps between the source and the invaded region, and the dynamics of demographic expansion after each introduction. Determining the routes of introduction — the geographic pathways of the propagules between the source and the invading populations — therefore provides not only information about the history of the invasion process, but also information about the origin and construction of the genetic composition of the invading populations (Dlugosch and Parker, 2008).

Two types of methods are traditionally used to make inferences about introduction routes: 1) Direct methods based on interception data and/or historical records of the presence or absence of the species (e.g. the study of Argentine ant by Suarez *et al.*, 2001), and 2) indirect methods based on genetic relationships between populations (e.g. Ciosi *et al.*, 2008; Kolbe *et al.*, 2004). Such indirect methods are usually based on the calculation of genetic distance (e.g. Goldstein *et al.*, 1999), assignment likelihood statistics (Rannala and Mountain, 1997) and parsimony networks (e.g. Voisin *et al.*, 2005).

One important limitation of the indirect methods is that they do not adequately take into account the demographic and genetic stochasticity of the history considered. The number of introduced individuals, their likelihood of becoming established and the time between introduction and demographic expansion may all be considered as random variables able to adopt various values. In addition, for a given set of demographic parameters, chance strongly affects the genetic composition of the samples studied because (i) genetic drift since the foundation event may have affected the genetic composition of the source population; (ii) genetic bottlenecks often occur during

68 the first few generations after introduction, due to the limited number of founders and the small size
69 of the newly founded population; (iii) mutational events may occur at any stage and (iv) limited
70 numbers of individuals are usually collected during the field sampling of populations for genetic
71 analysis. The level of stochasticity in introduction histories is therefore generally high, likely to have
72 profound consequences and may considerably decrease the validity of the results obtained by indirect
73 genetic methods.

74 A second general problem of these direct and indirect methods is that they do not allow
75 probabilistic estimations of competing introduction scenarios (but see Gaggiotti *et al.*, 2004). In
76 practice, an introduction scenario is chosen because the data are more consistent with that scenario
77 than with any other. It is, in a sense, a binary decisional process (accept or reject), in which the
78 relative likelihoods or weights of alternative scenarios are not known.

79 We evaluated here a model-based Bayesian method taking into account the stochasticity of the
80 demographic and genetic processes involved and making it possible to calculate the relative
81 probabilities of competing introduction scenarios. The Bayesian nature of this method makes it
82 possible to make use of prior historical, biological and genetic information about the system. It is
83 also based on a stochastic model linking demography and genetics — the coalescent model of the
84 genealogical process (Hudson, 1990; Kingman, 1982; Nordborg, 2001) — to provide a simple and
85 efficient population genetic model of drift and mutation particularly useful for handling complicated
86 evolutionary scenarios. Estimations associated with demographic and genetic models often imply a
87 full likelihood calculation, which is difficult for complex scenarios such as biological invasions.
88 Approximate Bayesian computation based on summary statistics (ABC, Beaumont *et al.*, 2002) does
89 not require likelihood computation, only require the ability to simulate datasets, and makes it
90 possible to handle large datasets, such as data for hundreds of individuals genotyped at tens of
91 microsatellite loci. This method was recently successfully used to estimate the demographic

parameters of various complex scenarios (Fagundes *et al.*, 2007; Pascual *et al.*, 2007) and for model selection (Cornuet *et al.*, 2008; Miller *et al.*, 2005; Pascual *et al.*, 2007).

We used simulated datasets to evaluate the utility of the ABC method for inferring the correct introduction routes of invasive populations. We considered microsatellite-like datasets, as this category of highly variable molecular markers have proved successful for addressing questions relating to invasive species (e.g. Estoup *et al.*, 2004; Fagundes *et al.*, 2007; Miller *et al.*, 2005; Pascual *et al.*, 2007). We evaluated this method in the simple but crucial case of two invasive populations strongly suspected to have originated from the same source population. The key issue is determining whether the invasive populations originated from two independent introduction events or serially, from a single introduction. This simple situation is the basis for retracing more complex multipopulational introduction histories (e.g. Miller *et al.*, 2005). We have illustrated certain features of our simulation-based study by applying the ABC method to a recently published dataset for invasive populations of the western corn rootworm, *Diabrotica vigifera virgifera* (Ciosi *et al.*, 2008; Kim and Sappington, 2005; Miller *et al.*, 2005)

MATERIALS AND METHODS

Models of introduction: We considered three models in which two invasive populations originate from the same source population. The dataset consist of genotypes at statistically independent microsatellite loci obtained from a sample of diploid individuals collected from the invasive and source populations. These populations may be related through three different scenarios (see Figure 1).

The independent introduction scenario: Both introduced populations were founded independently from the source population. We assume that at each introduction, the invading

117 population experienced a bottleneck because the number of founders is much lower than the number
118 of individuals in a stable population. In this first scenario, each introduced population experienced a
119 single bottleneck of possibly different intensities.

120 *The serial introduction scenario*: Only one introduced population originated from the source
121 population, with the second introduced population originating from the first. The first introduced
122 population thus experienced a single bottleneck (as in the independent introduction scenario), but the
123 second introduced population experienced two successive bottlenecks, the first being common to
124 both introduced populations.

125 *The unsampled population scenario*: Previous studies have shown that some invasive
126 populations may remain undetected but may play important role in the invasion dynamics of some
127 species (e.g. Roman, 2006; Saltonstall, 2002). In the two former scenarios, all the populations
128 concerned were sampled, but in the unsampled population scenario, the two invasive populations
129 were founded independently from an undetected and hence unsampled population, itself introduced
130 from the source (Figure 1). Each invasive population experienced two bottlenecks, the first being
131 common to both invasive populations. It is important to consider this scenario, because it
132 superficially resembles an "independent introduction scenario", with two introduced populations
133 independently founded from a common population, but it actually corresponds to a single
134 introduction from the native range. This is because only one introduced population, the unsampled
135 population, originated from the native population. The variability of alleles and genetic combinations
136 present in the invading populations is thus constrained by the genetic variation present in a single
137 introduced population, rather than in two independently introduced populations. The likelihood of
138 obtaining genetic combinations adapted to the new habitat is therefore lower than that in an
139 independent introduction scenario. The "unsampled population scenario" is hence both historically
140 and evolutionarily substantially different from the independent introduction scenario.

Historical and demographic parameters were the same for all introduction models. Each introduced population *i* was founded by individuals originating from its source population t_i generations ago and was characterized by an effective number of founders, Nf_i, remaining constant for a few generations (bottleneck duration *BDi*) and then instantaneously reaching a larger stable effective population size, N_i, (see bottom part of Figure 1). Parameters t_i, Nf_i and *BDi* can take distinct values in the various populations. In the meantime, the source population is assumed to have maintained a constant effective size N_s. The introduced populations are assumed to be isolated from each other and from the source population after the introduction, with no exchange of migrants. We also assume that no repeated introductions occurred at the same location.

ABC estimation of the posterior probabilities of scenarios: The posterior probabilities of competing scenarios were estimated with statistical methods implemented in the DIYABC program (Cornuet *et al.*, 2008) available from http://www.montpellier.inra.fr/CBGP/diyabc/. We simulated a large number (usually $3x10^5$) of genetic datasets under the coalescent model (Hudson, 1990; Kingman, 1982; Nordborg, 2001), using the three introduction scenarios according to their prior probability and their parameter values drawn from prior distributions. Each typical genetic dataset contained the diploid genotypes, at ten independent microsatellite loci, of 30 individuals sampled from each of the two invasive and the source populations. Diploid genotypes were generated assuming Hardy-Weinberg equilibrium by randomly grouping gene copies by pair. The summary statistics for each simulated dataset are recorded, together with the label of the scenario used for the simulation, in a file called the "reference table". As described by Beaumont *et al*. (2002), we calculated the Euclidean distances between each simulated dataset and the observed target dataset in the space of the summary statistics (standardized by the standard deviation of the simulated summary statistics) and these distances were then used to estimate the posterior probabilities of the scenarios.

As we used simulation to evaluate the ABC method, the target "observed datasets" were also simulated and will hereafter be referred to as "pseudo-observed datasets".

The posterior probabilities of the introduction scenarios were estimated by three methods. The first estimator, the "direct estimator", is the frequency of Euclidean distances associated with scenario *i* among distances below a specified threshold *z* (Cornuet *et al.*, 2008). The second estimator, derived from the k^{th} nearest-neighbor density estimator (equation 1.1 of Terrel and Scott, 1992), is called the "KN estimator" and is defined as:

$$P_{KNi} = \frac{\left(\frac{1}{\delta_k}\right)^d}{\sum_{i=1}^{n}\left(\frac{1}{\delta_k}\right)^d}, \qquad (1)$$

where δ_k is the k^{th} smallest Euclidean distance for the introduction scenario *i* and *d* is the number of statistics used to summarize the data. As explained by Fagundes et al. (2007), the third estimator, the "PL estimator", is based on the idea that we can sample the scenario indicator *i* (where *i* = 1...m for scenarios 1, ..., *m*) from its prior and treat it as a categorical random variable in the ABC simulations. We can then apply a categorical regression (a polychotomous logistic regression) and an Epanechnikov kernel, as described by Beaumont (2008) to estimate the posterior probability of scenario *i*. A proportion *y* of the data points, corresponding to the smallest Euclidean distances, was used in the regression. Confidence intervals of the PL estimator were computed as described in Cornuet *et al.* (2008).

The selection of suitable *z*, *k* and *y* thresholds is a difficult task, requiring cross validation procedures for each scenario tested. Low values for these thresholds result in the generation of accurate estimates, but high variances of the estimates. Large values may result in inaccurate estimates and low variances of the estimates. We chose to use the lowest values ensuring both a small variance of the probability estimate among 20 reference tables and 1,000 pseudo-observed

187 datasets and good stability — i.e. weak variation of the estimates with variations of *z*, *k*, and *y*
188 (results not shown). The results obtained for *z* and *k* values between 50 and 1000 and for *y* values
189 between 1000 and 100,000 (not shown) were all qualitatively similar to those presented below. A
190 value of 100 for *z* and *k*, and a value of 10,000 for *y* were hence selected for all calculations when not
191 indicated differently.
192
193 **Summary statistics:** Genetic variation was summarized within and between populations, using
194 the following statistics: the F_{ST}-values between pairs of populations (Weir and Cockerham, 1984)
195 and the mean individual assignment likelihoods of individuals collected in population *i* and assigned
196 to population *j* (Li(j Pascual *et al.*, 2007); the mean number of alleles per locus, the expected
197 heterozygosity (Nei, 1987), and the mean variance of the absolute allelic size (Estoup *et al.*, 2004)
198 computed for each population (*A*, *H*, *V*) or for each pair of populations (i.e. by pairwise pooling of
199 population samples) (*A2P*, *H2P* and *V2P*). The default set of statistics (hereafter referred to as
200 "default stat") was *A2P*, *H2P*, *V2P*, F_{ST}-values and $L_{i\to j}$ and hence consisted of 18 statistics. We also
201 used the set of statistics used by Miller *et al*. (2005) ("Miller stat": *A*, *H*, *M* (Garza and Williamson,
202 2001), F_{ST}-values, and $L_{i\to j}$, which also correspond to a total of 18 statistics) and by Beaumont
203 (2008) ("Beaumont stat": *A*, *H*, *V*, *A2P*, *H2P*, and *V2P*, for a total of 18 statistics). We compared the
204 results obtained with "default", "Miller" and "Beaumont stat" in order to select the best set of
205 summary statistics.
206
207 **Prior distributions of parameters:** We kept our simulation study as generic as possible in the
208 context of invasion biology. Each of the three competing scenario (independent, serial and
209 unsampled) was assumed to be equally probable. The default set of prior distributions of the
210 historical, demographic and mutational parameters is shown in Table 1. We used a generalized
211 stepwise mutation model (GSM) to simulate mutations at the molecular markers of interest (i.e.

212 microsatellites; (Estoup *et al.*, 2002)). A mean mutation rate across loci $\overline{\mu}$ was first drawn from its
213 distribution, then single locus mutation rates μ were drawn from a Gamma distribution with mean $\overline{\mu}$
214 and shape parameter 2 (rate=$2/\overline{\mu}$). For each locus, the coefficient *P* of the geometric distribution of
215 repeat units by which a new mutant allele differs from its ancestor was drawn from an exponential
216 distribution with mean of 0.22 (Miller *et al.*, 2005).
217
218 **Simulation of pseudo-observed datasets:** For each introduction scenario, 1,000 pseudo-
219 observed genetic datasets were simulated using the DIYABC program as for the calculations above.
220 The demographic, historical and mutational parameters used to simulate pseudo-observed data were
221 drawn from probability distributions with regions of positive probabilities (i.e. supports) included
222 within those of the prior distributions (Table 1). This assumption suggests that the knowledge about
223 the biological system studied is sufficient to correctly choose the prior distributions. This procedure
224 for simulating the test datasets was preferred over the more traditional strategy of fixing all but one
225 of the parameters and evaluating the effect of the unfixed parameter, because the demographic,
226 historical and mutational parameters were thought likely to act together to produce the pseudo-
227 observed statistics and inference results. We used specific statistical treatments based on linear
228 models to analyze the effect on posterior probabilities of varying the parameters used to simulate
229 pseudo-observed datasets (see Supplementary information for details). For these analyses we used a
230 specific set of prior distributions: the "alternative prior distributions" described in Table 1.
231
232 **Performance of the ABC approach:** The selected scenario was defined as the most probable
233 of the scenarios considered. The performance of the ABC method was evaluated by measuring its
234 accuracy. The accuracy of a classification method is the frequency at which it correctly selects the
235 "true" introduction scenario from among the tested scenarios. The term accuracy is used in the
236 following with this meaning only. For each estimator, we also calculated the area under the curve

(AUC) of a receiver operating characteristics (ROC) graph (Fawcett, 2006). The largest AUC corresponds to the method giving the best compromise between true positive and false negative rates.

We evaluated the effect of the number of loci (5, 10, 20 and 50), sample size (15, 30 and 60 diploid individuals), and the number of simulated datasets in the reference table ($3x10^4$, $3x10^5$ and $3x10^6$). The performance of the ABC method was also compared with that of indirect methods based on raw F_{ST} or $L_{i\rightarrow j}$ statistics. With such methods, it is possible to infer introduction scenarios based on the following rules: the source of each introduced population is the sample for which the $L_{i\rightarrow j}$ value is the largest or the F_{ST}-value the smallest. Considering populations S, 1, and 2 in Figure 1, the independent introduction scenario (S→1, S→2) is expected to produce the following relationships: $L_{1\rightarrow S} > L_{1\rightarrow 2}$, $L_{2\rightarrow S} > L_{2\rightarrow 1}$, and $F_{ST12} > F_{STS1}$, $F_{ST12} > F_{STS2}$. The serial introduction scenario (S→1→2) is expected to produce the following relationships: $L_{2\rightarrow 1} > L_{2\rightarrow S}$, $L_{1\rightarrow S} > L_{1\rightarrow 2}$, and $F_{STS2} > F_{ST12}$, $F_{STS2} > F_{STS1}$. No specific hierarchical relationship is expected for the unsampled population scenario.

Robustness of inferences: The robustness of the ABC method was evaluated by analyzing the effect of various erroneous beliefs on the biological system studied. Such errors were investigated by altering either the scenario or the distributions of parameters used to simulate the pseudo-observed datasets (true introduction scenario and true parameter distributions) whereas prior parameter distributions and the three competing scenarios (independent, serial and unsampled) used for the inference remained unmodified. The following list of tests presents the various modifications of the true scenario and the true parameter distributions.

Test 1: The supports (i.e. the regions of positive probabilities) of the true parameter distributions used to generate the pseudo-observed datasets were not included in those of the prior distributions (Table 1).

Test 2: False source in the invaded area: In the true introduction scenario, an introduced unsampled population has generated two serially introduced populations (unsampled serial introduction scenario in Figure 2).

Test 3: False source in the native area. In the true introduction scenario, the sampled population in the native area had actually diverged from the real source population $t_{\text{false source}}$ generations ago, without a bottleneck (false source independent introduction scenario in Figure 2).

Test 4: Two sources. In the true introduction scenario, the two invasive populations were founded from two sources that had diverged t_{sources} generations ago, and the source of only one invading population was sampled (two sources introduction scenario in Figure 2).

Test 5: False sequence of introductions. The sequence of introduction events was inverted due to uncertain, if not erroneous information concerning the dates on which the species was first sighted. The true inverted serial introduction scenario (Figure 2) was hence source→pop2→pop1 whereas the tested serial scenario was source→pop1→pop2. In the same vein, the true inverted independent introduction scenario was source→pop2 and source→pop1 whereas the tested serial scenario was source→pop1 and source→pop2.

The prior distributions for the parameters used for tests 2, 3, 4 and 5 were those of the default set (Table 1), whereas $t_{\text{false source}}$ and t_{sources} were drawn from Uniform[40;500].

Application to the western corn rootworm, a pest beetle invading Europe: The western corn rootworm (WCR), *Diabrotica virgifera virgifera* Leconte, is a univoltine chrysomelid and is a major pest of corn in North America. It has recently been introduced into Europe. Several disconnected invading populations have been observed in Europe, including two large expanding outbreaks first observed in Serbia in 1992 (hereafter referred to as the Central European outbreak), and in North Western (NW) Italy in 2000. Miller *et al*. (2005) studied the introduction routes of WCR, using a previous version of the ABC method (using the KN and direct estimators) and

concluded that these two large expanding outbreaks were independently founded by individuals originating from North America. Here, we check this result using the revised ABC method and, most importantly, we aim to illustrate the effect of considering erroneous source populations. We performed an ABC analysis on the NW Italian and Central European outbreaks which were sampled in 2003 and described by Miller *et al*. (2005), using various North American samples as putative sources (samples collected in Texas and Pennsylvania in 2004, in Kansas, Nebraska, Iowa, Ohio, Illinois, Delaware and New York in 2003 (Kim and Sappington, 2005), and in Arizona in 1998 (Ciosi *et al.*, 2008)). These American samples displayed low to medium levels of genetic differentiation (F_{ST}-values ranging from 0 within the Corn Belt to about 0.06 between Delaware and Arizona). We used the genotypes obtained at eight microsatellite loci (Miller *et al.*, 2005).

The default set of priors detailed in table 1 was used, with the following minor modifications: the years of introduction of WCR in Europe were drawn from uniform distributions bounded by 1986 and 1991 for Central Europe, and by 1995 and 1999 for the NW Italian outbreak. The generation time for *D. virgifera* is one year. Given the available data on the fertility and population growth rates of WCR (Toepfer and Kuhlmann, 2005), we uniformly draw duration of bottlenecks following the introduction between 1 and 5 years, as in the study by Miller *et al.* (2005).

RESULTS

Inferring introduction scenarios: We found the ABC method for inferring the introduction routes of invading populations to be efficient in most of the simulations studied whatever the estimator of the posterior probabilities (Table 2). When scenario selection was based on the highest posterior probability value, between 89 and 96 % of the pseudo-observed datasets were correctly assigned to the true scenario (mean accuracy of 0.92). Note that with this classification criterion, a

310 classification can be correct although the estimated probability of the true scenario is below 0.9. This
311 explains why the frequency of correct classification (*f* in Table 2) is generally high although
312 $f(Pi>0.9)$ may be low. The unsampled population scenario was well recovered by the ABC method.
313 The traditional indirect method (described in the "Performance of the ABC approach" section),
314 in which raw F_{ST}-values are used to infer the introduction scenario, correctly classifies the pseudo-
315 observed datasets to the independent or serial scenario (Table 2). It is worth stressing, however, that
316 this method cannot correctly classify the unsampled population scenario and selects the independent
317 or serial introduction scenario for 66 % of the datasets simulated under the unsampled population
318 scenario. The method based on mean individual assignment likelihood did not provide satisfactory
319 results. More than 60% of the pseudo-observed data simulated under the serial introduction scenario
320 could not be assigned to either the independent or serial introduction scenario. Like the F_{ST}-based
321 method, the assignment likelihood method cannot classify the unsampled population scenario. It is
322 noteworthy that when an introduction is incorrectly selected, the differences between assignment
323 likelihoods or between F_{ST} values whose relationships are used to infer the scenario are lower than
324 when the scenario is correctly selected. For instance the differences $L_{1\to S} - L_{1\to 2}$, $L_{2\to S} - L_{2\to 1}$, $F_{ST12} - F_{ST S1}$, $F_{ST12} - F_{ST S2}$ are larger if the independent scenario is correctly selected than if it is incorrectly
325
326 selected (e.g. when the unsampled population scenario is true).
327 Additional analyses showed that when the bottleneck population sizes used to simulate the
328 pseudo-observed datasets were larger than in the default conditions, the simple indirect methods
329 performed much less well than the ABC method. With bottleneck population sizes drawn from a
330 uniform prior distribution between 50 and 500 and from a uniform distribution to generate the
331 pseudo-observed datasets bounded by 100 and 500 (all other distribution being the same as before),
332 we obtained the following frequencies of correct scenario identification: 67.4% for the F_{ST} method,
333 70% for $L_{i\to j}$ and 83.6% for ABC when the true scenario was the independent scenario; and 65.3%
334 for the F_{ST} method, 33.1% for $L_{i\to j}$, and 80.4% for ABC when the serial scenario was true. As

335 already mentioned, the unsampled population scenario could not be identified by the F_{ST} and $L_{i\to j}$
336 methods. It was correctly identified by the ABC method at a rate of 53.9%.
337
338 **Comparing estimators of introduction scenario probability:** The expected values of the
339 posterior probabilities of our complex competing scenarios are unknown, but a 'good' estimator
340 should frequently assign the pseudo-observed datasets to the introduction scenario used to simulate
341 them. We found that the three estimators behaved very similarly in this respect (Table 2).
342 Another desirable property of the posterior probability estimator is its small variance among
343 independent analyses of the same data. The standard error of the posterior probability estimator was
344 calculated by analyzing the 3,000 simulated pseudo-observed datasets 30 times, with 30 independent
345 reference tables. The PL estimator had a standard error between reference tables one half to one third
346 that of the other estimators (see the values in brackets in Table 2).
347 Importantly, as mentioned in the Materials and Methods section, the PL estimator displayed a
348 remarkable stability — i.e. weak variation of the estimates with variations of *y*, the proportion of the
349 data points used in the regression. We observed almost no variation of PL estimates with *y* values
350 varying between 1000 and 100,000.
351 Finally, we could calculate confidence intervals (CIs) for the posterior probability of the PL
352 estimator. Use of the lower limit of the 95% CI of the maximal probability scenario being greater
353 than the upper limit of the 95% CI of the other scenarios as the classification criterion gave an
354 accuracy similar to that obtained with the largest posterior probability (Table 2). The overall
355 properties of the PL estimator were thus considered better than those of the direct and KN estimators.
356 We therefore present only the results obtained with the PL estimator below.
357

Effects of model parameter values on inferences: The alternative parameter distributions and alternative prior distributions (Table 1) used to address this particular question provided good discrimination results: AUC = 0.99, accuracy = 0.95.

We performed a statistical analysis of the effects of variation in the parameters on the posterior probabilities of scenarios, using a linear model-based approach. The results of the linear model fits are shown in Supplementary Table 1. Linear models accounted for 22%, 31%, and 45% of the variance when the true scenario was the independent, serial or unsampled population scenario, respectively. Most significant and larger effects dealt with bottleneck parameters: the posterior probabilities of the independent and serial scenario, when true, were negatively correlated with the intensity of drift in the two introduced populations during bottleneck events. The probability of the true scenario was higher for larger bottleneck population sizes and shorter bottleneck durations. When the unsampled population scenario was the true scenario, the effect of drift intensity in the introduced populations depended on the population considered. Drift intensity had a negative effect in the more recent introduced population (population 2 in Supplementary Table 1), and a variable effect in the older introduced population (population 1) and in the unsampled population. Increasing the number of effective founders in the more recently introduced and unsampled populations had a positive effect on the probability of the true scenario below a certain population size and a negative effect above this threshold. Parameters affecting genetic diversity in the source population (i.e. N_{source} and $\bar{\mu}$) had a positive effect on the posterior probability of the independent and serial introduction route scenarios, but no significant effect for the unsampled population scenario. The dates of introductions and stable effective population sizes after the bottleneck period, for introduced populations, had no effect on the posterior probability for any of the scenarios.

Influence of ABC summary statistics, number of individuals, loci and reference table size: The number of datasets in the reference table, the number of diploid individuals sampled per

population and the number of loci have a positive effect on the AUC, the probability of the true scenario and the rate of identification of the true scenario (Table 3). The ABC method was found to be sensitive to the number of loci used, but provided good results with as few as five loci (more than 70 % of scenarios correctly identified). Reducing the sampling effort to 15 individuals per sample decreased the accuracy by about 5 percentage points with respect to the default sample size (30 individuals). Doubling the default sample size per population did not increase accuracy. The use of $3x10^6$ datasets per reference table gave no improvement in the correct classification rate over the use of $3x10^5$ datasets, and $3x10^4$ datasets appear sufficient to reach a similar accuracy. Finally, all tested sets of ABC summary statistics ("default", "Miller", "Beaumont stat") gave similar results, with an accuracy of between 91 and 92 % (Table 3).

Robustness of inference and effect of erroneous prior beliefs:

Test 1: Errors in demographic, historical and genetic parameter priors (Table 4). Using priors for *Ns* with larger values than that of *Ns* used to simulate pseudo-observed datasets decreases the posterior probabilities of the true scenarios and increases their variances. Consequently, the correct introduction scenario is recovered less often, even if it is correctly inferred in more than 70% of the simulated cases. Conversely, if the priors used for Nf_i are too small, the rate of correct classification increases markedly when the true scenario is the independent or serial scenario. The true unsampled population scenario is poorly recovered, however, and the distribution of the posterior probabilities of the three scenarios is generally flat in this case (less than 7% of the probabilities exceed 0.9, versus 47% for the default parameters; Table 2).

If bottleneck duration priors are too small, scenario identification becomes much less accurate, particularly if the serial scenario is true. In this situation, the unsampled population scenario is erroneously selected in 72 % of the simulated cases. However, as previously stated, the distributions

of the posterior probabilities of the three scenarios are flat (less than 10% of the probabilities exceed 0.9, versus more than 30% for the default parameters; Table 2).

Erroneous beliefs concerning mutational parameters had limited effects on inference accuracy, unless the mean mutation priors are too large.

A large error in the timing of introduction events (such as the actual introductions taking place about 100 generations before they were believed to occur) had only a small effect on inferences. However, the unsampled population scenario tended to be selected more frequently when the serial introduction scenario was true than in the default situation (26% versus 9%).

Test 2: False source in the invaded area. The ABC method selects the serial scenario when the unsampled-serial scenario is true with about the same frequency as when the serial scenario is true.

Test 3: False source in the native area. When the two populations were introduced independently only 39% of the simulated cases were assigned to the independent scenario and 61 % falsely assigned to the unsampled population scenario. This effect depends on the level of differentiation between the actual and the believed source (see Supplementary Figure 1). Moderate to high levels of divergence ($F_{ST} > 0.02$) led to selection of the (false) unsampled population scenario. However, if only independent and serial scenarios are considered, the rate of recovery of the true scenario is high (i.e. 84.8% for independent and 100% for serial scenarios; results not shown).

Test 4: Two sources. If the two invasive populations originate from two diverging sources (quartiles of F_{ST} distribution: 0.008, 0.014, 0.021) but it is believed that one of these sources is the source of both populations (see “Two sources” in Table 4), the independent scenario is recovered as frequently as in the default situation.

Test 5: False sequence of introductions. When the true serial scenario is source→pop2→pop1 but the tested scenario is source→pop1→pop2, the unsampled population scenario is incorrectly selected in most cases (94%). However, no effect on posterior probabilities was observed when the independent or the unsampled population scenario was the true scenario. This led us to consider this "inverse serial scenario" to the scenarios tested, resulting in the consideration of four, rather than three competing scenarios (i.e. independent, serial, inverse serial and unsampled scenarios) in a dedicated analysis (Supplementary Table 2). The true inverse serial scenario was correctly recovered in 91 % of the simulated cases (Supplementary Table 2). Adding an alternative serial scenario to the set of scenarios tested only slightly decreased the global accuracy of the ABC method (from 92 to 89%). For the sake of clarity we considered this fourth (inverse serial) scenario in this specific analysis (Supplementary Table 2) only.

Application to the western corn rootworm: We performed ten successive ABC analyses of introduction scenarios, using each of the ten North American samples as the source of the European populations. The serial scenario had a null posterior probability whatever the source population used. If one of the samples collected from the central and eastern part of the USA (from Kansas to the state of New York) was used as the source, the probability of the independent introduction scenario was high (between 0.7 and 0.9 with narrow 95 % confidence intervals), except for the sample from Illinois, which gave a probability of 0.46 (95 % CI: 0.39 - 0.53). Interestingly enough, we found that the unsampled population scenario was selected with probabilities of 1 (95% CI: 1-1) and 0.88 (95 % CI: 0.85-0.91) if we used the samples from Arizona and Texas, respectively, as the source.

For the sake of simplicity, we defined the "true" source population of the European outbreaks as the North American population with the lowest F_{ST}-value with respect to the NW Italian and Central European samples. The sample collected in Delaware was thus considered the most probable source of both European populations. Figure 3 shows the relationships between the posterior

probability of the independent scenario and the pairwise F_{ST}-values obtained by comparison between each "false" source population used to compute the probability and the Delaware sample. Consistent with our previous simulation results, there was a marked inverse correlation between the probability of the independent scenario and F_{ST}. Figure 3 also shows that when only two scenarios are tested (independent and serial introduction scenarios), the posterior probability of the independent scenario is 1, with a small 95% CI for all of the US samples considered as the source except for the Arizona sample. The ABC method correctly selects the unsampled population scenario, when true, whether or not beliefs concerning the source in the area of origin are erroneous (Table 4). Thus, our results for WCR strongly support the hypothesis of an independent introduction scenario with a source population within or close to the central or eastern part of the USA. They also nicely illustrate the potential bias that is likely to arise when considering a wrong genetically differentiated source population.

DISCUSSION

We evaluated the ability of an ABC method to infer introduction routes in the context of biological invasions. Using a simulation-based approach, we demonstrated that, when two invading populations and a source population are considered, the ABC method very frequently identifies the "true" introduction scenario. This method also provides an estimate of the posterior probability of the competing scenarios and confidence intervals for each probability. It is worth stressing that a set of n invasive populations with a common source population in the native area can be analyzed by studying $n(n\text{-}1)/2$ pairs of invasive populations with the independent, serial and unsampled population scenarios (Miller *et al.*, 2005).

Calculation of the posterior probabilities for each competing scenario is of key importance, as it makes it possible to quantify our level of confidence in the choice of a specific scenario (as

opposed to the others). The Bayesian nature of the ABC method confers on this method the advantage that prior information can be incorporated such that the competing scenarios have different weights. For instance, the dates of the first observations of introduced populations can be used to weight the serial and inverse serial introduction scenarios, as proposed by Miller *et al*. (2005).

Methods based on F_{ST} or $L_{i\to j}$ values to which the ABC approach was compared are not the most powerful (as compared, for example, with full-likelihood methods). However, "distance" methods, some of which are tree-based, are commonly used in invasion biology to determine the most probable sources of invasive populations and, by extension, to determine introduction routes of invasive populations. We found that F_{ST} or $L_{i\to j}$ methods were less reliable than the ABC method for at least three reasons: (1) they cannot include the unsampled population scenario among the tested models; more generally, as soon as a tested scenario is not directly translatable into a simple hierarchy between F_{ST}-values or between $L_{i\to j}$ values, it cannot be evaluated. More importantly, when the unsampled population scenario is true, these methods erroneously identify either the independent or the serial scenario as the true scenario; (2) they result in incorrect classification when bottleneck intensities are moderate and, most importantly, (3) they do not provide probabilities for the introduction routes tested or any other measurement of confidence in the choice made. Many previous studies in the field of invasion biology (e.g. Kolbe *et al.*, 2004; Voisin *et al.*, 2005) have been based on methods using distance trees or parsimonious networks built from nucleotide sequence data (often mitochondrial or chloroplast DNA). Drawbacks (1) and (3) of the F_{ST} and $L_{i\to j}$ approaches probably also apply to these methods. A simulation-based study is required to evaluate drawback (2).

We compared three different estimators of posterior probabilities: the direct estimator, the k^{th} neighbor density estimator (KN) and the polychotomous logistic regression estimator (PL). We found that the PL estimator had desirable properties, such as a low sensitivity to the choice of threshold and a low variance. However, it often provided lower values for true scenario probabilities

than the KN estimator. The reanalysis in this paper of some of the data of Miller *et al*. (2005) using the PL estimator confirmed, with high levels of confidence, that the North Western Italian and Central European outbreaks of WCR resulted from independent introductions from North America (most probably from the central north-eastern part of the USA).

The default parameters used in our simulation settings correspond to an *a priori* unfavorable situation in which the invasions were recent. In such conditions genetic differentiation would be expected to be minimal and signatures of the introduction history due to long divergence times between populations should be absent. Consistent with these assumptions, we showed that bottleneck intensities were key factors determining the posterior probabilities of introduction routes. We might initially have expected stronger bottleneck intensities to favor scenario discrimination, because drift pulses during bottlenecks are at least partly responsible for the genetic signature of introductions. However, we found the opposite to be true. Accuracy was negatively affected by bottleneck intensity, probably because intense bottlenecks (particularly during the first introduction event) tend to generate patterns expected under the unsampled population scenario, in which the gene genealogies of both introduced populations suffer from two successive bottlenecks.

We found that our approach was robust to many types of error in prior beliefs. In particular, errors concerning mutational parameters and the dates of events were found to have negligible effects on classification results. If only the independent and serial scenarios are considered, the ABC method is almost insensitive to error concerning demographic parameters, including bottleneck intensity. The unsampled population scenario may increase the susceptibility of the analysis to errors in prior beliefs, but including this scenario in the set of tested scenarios is crucial, as it avoids confusion between multiple and single introduction scenarios. It might be worth reanalyzing some of the previously reported descriptions of multiple introduction scenarios that may actually be the result of unsampled population scenarios.

We found that in case of geographic genetic structure within the native area, an error concerning the source population in this area resulted in misleading results being obtained with the ABC method. The probability of the unsampled population scenario increases with the level of genetic differentiation between the true and false sources when the true scenario is an independent introduction scenario. We have illustrated this effect with real datasets from the western corn rootworm invasion in Europe. The North Western Italian and Central European outbreaks of WCR probably resulted from independent introductions of individuals originating in the northern US (Ciosi *et al.*, 2008; Miller *et al.*, 2005). If we considered the source population to have originated from the central or north-eastern USA, the independent scenario was selected with considerable support. However, if we considered samples from Texas or Arizona, genetically differentiated from the samples collected in central or north-eastern USA, to be the source, the unsampled population scenario was selected with high posterior probabilities. Again, if the unsampled population scenario is not considered among the tested scenarios, then the ABC method is insensitive to errors concerning the source population. We also demonstrated that in cases of uncertainty concerning the order of introduction events, a satisfactory solution is to include the inverse serial scenario among the tested scenarios (as shown by Miller *et al.*, 2005). Our study shows that this approach prevents the misclassification of serial introduction scenarios as unsampled population scenarios, with no major loss in classification accuracy.

We found that posterior probability distributions were often flat when errors of classification were observed, resulting in large 95% CI and low levels of confidence in the results obtained. Such flat posterior probabilities should therefore be interpreted with caution, as they may indicate errors in parameter priors and/or model specification. The comparison between observed summary statistics and simulated statistic distributions may also be used to detect such errors (Pascual *et al.*, 2007). In To summarize this part on the effect of erroneous beliefs, we would suggest the reader to 1) choose broad support of priors (i.e. region of the prior with positive probabilities) to ensure that it includes

556 the “true values” of parameters, 2) be sceptical when obtaining low maximal probabilities (less than
557 0.7) and modify the prior distribution and/or model specification, and 3) compare the results
558 obtained with and without the unsampled and the inverse serial scenarios to be able to detect error in
559 scenario specifications.

560 The evaluation of the ABC method presented here was subject to several limitations. This
561 study deals with a simple biological invasion situation in which only two invading populations and a
562 single source are considered. Although this setting can be used as the basis for retracing more
563 complex multipopulational introduction histories (e.g. Miller *et al.*, 2005), the ABC method
564 implemented in DIYABC can handle complex scenarios of introduction routes with a large number
565 of populations directly (see Cornuet *et al.* (2008) for methodological details and illustrations). For
566 instance, admixture can be modeled in DIYABC and the assumption of a single source can easily be
567 relaxed by considering two populations that diverged some generations ago and are the sources of
568 the introduced populations. We also assumed that there was no migration between invading
569 populations and no recurrent introductions into each invading population. Although considering
570 migration should be useful in certain cases, the absence of migration may be a reasonable
571 approximation in many circumstances, particularly when the invading populations are geographically
572 distant from the original source population and from each other (as for the Italian and Central
573 European outbreaks of WCR or the populations of Drosophila subobscura introduced into North and
574 South America, Pascual *et al.*, 2007). Recurrent introductions from the same source population into
575 an invading population, although not probable in the case of recent introductions, would probably be
576 equivalent to increasing propagule size and would thus not be genetically distinguishable from a
577 single introduction with a larger propagule size (although this remains to be carefully tested). Finally,
578 it would be worth evaluating the ABC method for estimating probabilities of introduction routes
579 when other genetic markers such as DNA sequences, SNPs or AFLP markers are used.
580

581

582 ACKNOWLEDGMENTS

583

584 We would like to thank Christian Robert and Jean-Michel Marin, who helped us to compare
585 posterior probability estimators, Filipe Santos and Sylvain Piry for assistance with computer cluster
586 installation and Aurélie Blin and Pascal Chavigny for technical assistance. This work was funded by
587 the French National Research Agency (ANR) through project grants ANR-06-BDIV-008-01 to TG
588 and AE and ANR-BLAN-0196-01 to J-MC and AE.

589

590 Supplementary information is available at Heredity's website.

591

592

LITERATURE CITED

Beaumont M (2008). Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura S, Forster P and Renfrew C (eds) *Simulations, Genetics and Human Prehistory*. McDonald Institute for Archaeological Research: Cambridge, pp 135-154.

Beaumont MA, Zhang WY, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* **162:** 2025-2035.

Ciosi M, Miller NJ, Kim KS, Giordano R, Estoup A, Guillemaud T (2008). Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions and various patterns of reduced genetic diversity. *Molecular Ecology* **17:** 3614-3627.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ *et al.* (2008). Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics* **24:** 2713-2719.

Dlugosch KM, Parker IM (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology* **17:** 431-449.

Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet JM (2004). Genetic analysis of complex demographic scenarios: Spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution* **58:** 2021-2036.

Estoup A, Jarne P, Cornuet JM (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11:** 1591-1604.

Fagundes NJR, Ray N, Beaumont MA, Neuenschwander S, Salzano F, Bonatto SL *et al.* (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 17614-17619.

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27:** 861-874.

Gaggiotti OE, Brooks SP, Amos W, Harwood J (2004). Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology* **13:** 811-825.

Garza JC, Williamson EG (2001). Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10:** 305-318.

Goldstein DB, Roemer GW, Smith DA, Reich DE, Bergman A, Wayne RK (1999). The use of microsatellite variation to infer population structure and demographic history in a natural model system. *Genetics* **151:** 797-801.

Hudson RR (1990). Gene genealogies under the coalescent process. In: Futuyma DJ and Antonovics J (eds) *Oxford Survey in Evolutionary Biology*. Oxford University Press: Oxford. Vol. 7, pp 1-44.

640 Kim KS, Sappington TW (2005). Genetic structuring of western corn rootworm (Coleoptera :
641 Chrysomelidae) populations in the United States based on microsatellite loci analysis. *Environmental*
642 *Entomology* **34:** 494-503.
643
644 Kingman JFC (1982). The coalescent. *Stochastic Process Appl* **13:** 235-248.
645
646 Kolbe JJ, Glor RE, Schettino LRG, Lara AC, Larson A, Losos JB (2004). Genetic variation increases
647 during biological invasion by a Cuban lizard. *Nature* **431:** 177-181.
648
649 Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, Derridj S *et al.* (2005). Multiple transatlantic
650 introductions of the western corn rootworm. *Science* **310:** 992-992.
651
652 Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.
653
654 Nordborg M (2001). Coalescent Theory. In: Balding DJ, Bishop M and Cannings C (eds) *Handbook*
655 *of Statitical Genetics*. John Wiley & Sons, Ltd.: Chichester, England, pp 179-212.
656
657 Pascual M, Chapuis MP, Mestres F, Balanya J, Huey RB, Gilchrist GW *et al.* (2007). Introduction
658 history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC
659 methods. *Molecular Ecology* **16:** 3069-3083.
660
661 Rannala B, Mountain JL (1997). Detecting immigration by using multilocus genotypes. *Proceedings*
662 *of the National Academy of Sciences of the United States of America* **94:** 9197-9201.
663
664 Roman J (2006). Diluting the founder effect: cryptic invasions expand a marine invader's range.
665 *Proceedings of the Royal Society B-Biological Sciences* **273:** 2453-2459.
666
667 Saltonstall K (2002). Cryptic invasion by a non-native genotype of the common reed, Phragmites
668 australis, into North America. *Proceedings of the National Academy of Sciences of the United States*
669 *of America* **99:** 2445-2449.
670
671 Suarez AV, Holway DA, Case TJ (2001). Patterns of spread in biological invasions dominated by
672 long-distance jump dispersal: Insights from Argentine ants. *Proceedings of the National Academy of*
673 *Sciences of the United States of America* **98:** 1095-1100.
674
675 Terrel GR, Scott DW (1992). Variable kernel density estimation. *The Annals of Statistics* **20:** 1236-
676 1265.
677
678 Toepfer S, Kuhlmann U (2005). Natural mortality factors acting on western corn rootworm
679 populations: a comparison between the United States and Central Europe. In: Vidal S, Kuhlmann U
680 and Edwards CR (eds) *Western corn rootworm: ecology and management*. CABI Publishing:
681 Wallingford, UK, pp 95-119.
682
683 Voisin M, Engel CR, Viard F (2005). Differential shuffling of native genetic diversity across
684 introduced regions in a brown alga: Aquaculture vs. maritime traffic effects. *Proceedings of the*
685 *National Academy of Sciences of the United States of America* **102:** 5432-5437.
686
687 Weir BS, Cockerham C (1984). Estimating *F*-statistics for the analysis of population structure.
688 *Evolution* **38:** 1358-1370.
689

TABLE 1: Distributions of parameters

	Prior for the reference table		*Distribution for the pseudo-observed datasets*		
Parameters	*Default*	*Alternative*	*Default*	*Alternative*	*Analysis of erroneous prior specifications*
N_s and N_i	Uniform[1000; 20000]	Uniform[20000; 2000000]	Uniform[5000; 10000]	Uniform[50000; 1000000]	Uniform[500; 1000] (*Too large N_s values in prior*) Uniform[20000; 30000] (*Too low N_s values in prior)*
Nf_i	Uniform[1; 100]	Uniform[1; 300]	Uniform[5; 50]	Uniform[5; 200]	Uniform[100; 150] (*Too low N_f values in prior*)
BDi	Uniform[1; 10]	Uniform[3; 12]	Uniform[3; 7]	Uniform[5; 10]	Uniform[11; 15] (*Too low DB values in prior*)
$\overline{\mu}$	Uniform[10^{-4}; 10^{-3}]	Uniform[$8x10^{-5}$; $2x10^{-3}$]	$5x10^{-4}$	Uniform[10^{-4}; 10^{-3}]	$5x10^{-3}$ (*Too low $\overline{\mu}$ values in prior*) $5x10^{-5}$ (*Too large $\overline{\mu}$ values in prior*)
μ	Gamma(2; 2/$\overline{\mu}$)	Gamma(2; 2/$\overline{\mu}$)	Gamma(2; 2/$\overline{\mu}$)	Gamma(2; 2/$\overline{\mu}$)	
$\overline{P}$	0.22	Uniform[0.08; 0.36]	0.22	Uniform[0.10; 0.34]	
P	Exp($\overline{P}$)	Exp($\overline{P}$)	Exp($\overline{P}$)	Exp($\overline{P}$)	Exp(0.11) or $\overline{P}$ =0 (*Too large P values in prior*) Exp(0.44) (*Too low P values in prior*)
t_2	Uniform[10; 19]	Uniform[11; 50]	Uniform[13; 17]	Uniform[13; 47]	Uniform[113; 117] (*Too low Hist values prior*)
t_1	Uniform[20; 29]	Uniform[51; 90]	Uniform[23; 27]	Uniform[53; 87]	Uniform[123; 127] (*Too low Hist values in prior*)
$t_{unsampled}$	Uniform[30; 39]	Uniform[91; 120]	Uniform[33; 37]	Uniform[93; 117]	Uniform[133; 137] (*Too low Hist values in prior*)

Ns is the effective population size of the source, Nf_i, the effective number of funders, BD_i, the bottleneck duration, N_i, the effective size after growth of outbreak i (with i = 1, 2 or unsampled). Outbreak i was founded t_i generations before present. μ is the single locus mutation rates with mean $\overline{\mu}$ and P is the coefficient of the geometric distribution of repeat units with mean $\overline{P}$. Exp: exponential distribution. Alternative distributions were used to assess the effect of model parameter values on inferences (see text for details).

TABLE 2: Performance of the ABC and summary statistic-based methods

		Summary statistic methods		ABC method									
		f		f			f_{lim}	$\overline{P}$ (sd)			$f(Pi>0.9)$		
True scenario	Tested scenarios	F_{ST}	Li→j	Direct	KN	PL	PL	Direct	KN	PL	Direct	KN	PL
Independent	Independent	**0.97**	**0.98**	**0.92**	**0.92**	**0.89**	**0.88**	**0.77 (0.19) [0.039]**	**0.86 (0.21) [0.027]**	**0.81 (0.22) [0.016]**	**0.26**	**0.65**	**0.50**
	Serial	0.01	0.01	0.00	0.00	0.00	0.00	0.02 (0.05) [0.009]	0.01 (0.05) [0.003]	0.01 (0.03) [0.001]	0	0	0
	Unsampled	NC	NC	0.08	0.08	0.11	0.09	0.21 (0.17) [0.038]	0.13 (0.20) [0.027]	0.18 (0.21) [0.016]	0	0.10	0.10
Serial	Independent	0.03	0.00	0.00	0.00	0.00	0.00	0.00 (0.03) [0.003]	0.00 (0.03) [0.001]	0.00 (0.02) [0.000]	0	0	0
	Serial	**0.94**	**0.37**	**0.91**	**0.92**	**0.91**	**0.90**	**0.72 (0.15) [0.044]**	**0.81 (0.18) [0.045]**	**0.79 (0.17) [0.016]**	**0.08**	**0.45**	**0.34**
	Unsampled	NC	NC	0.08	0.08	0.09	0.06	0.28 (0.14) [0.044]	0.18 (0.18) [0.045]	0.21 (0.17) [0.016]	0	0	0
Unsampled	Independent	0.37	0.24	0.01	0.01	0.01	0.01	0.04 (0.10) [0.011]	0.03 (0.10) [0.006]	0.02 (0.08) [0.002]	0	0	0
	Serial	0.29	0.15	0.05	0.06	0.03	0.03	0.17 (0.16) [0.033]	0.10 (0.18) [0.026]	0.14 (0.14) [0.011]	0	0	0
	Unsampled	NC	NC	**0.93**	**0.93**	**0.96**	**0.94**	**0.79 (0.16) [0.039]**	**0.87 (0.19) [0.032]**	**0.84 (0.15) [0.013]**	**0.28**	**0.66**	**0.47**
Mean Accuracy		0.64	0.45	0.92	0.92	0.92	0.91						

Direct. KN, and PL represent three different estimation methods of posterior probabilities of introduction scenarios as described in the Matherials and Methods section. Indirect methods are based on raw F_{ST} or assignment likelihood Li→j values. f is the proportion of the simulated data classified into each tested scenario. The frequency at which the lower limit of the 95% confidence interval of the posterior probability of each scenario exceeds the upper limit of alternative scenarios is f_{lim} (computation possible only for the PL estimator). For the ABC method, the mean ($\overline{P}$) and the standard deviation (sd) of the posterior probability of each tested scenario and the frequency of posterior probabilities exceeding 0.9 ($f(Pi>0.9)$) are shown. In addition, the mean standard error of the posterior probability of each tested scenario among 30 reference tables is given in square brackets. NC: Not computable.

TABLE 3: Effect of number of loci and individuals per sample, reference table size and ABC summary statistics

				True introduction scenario											
				Independent				Serial				Unsampled			
Parameters		AUC	Accuracy	$\overline{P}_{indep}$ (sd)	f_{indep}	f_{serial}	$f_{unsampled}$	$\overline{P}_{serial}$ (sd)	f_{indep}	f_{serial}	$f_{unsampled}$	$\overline{P}_{ghost}$ (sd)	f_{indep}	f_{serial}	$f_{unsampled}$
Standard		0.991	0.92	0.81 (0.22)	0.89	0.00	0.11	0.79 (0.17)	0.00	0.91	0.09	0.84 (0.15)	0.01	0.03	0.96
No. loci	5	0.967	0.85	0.73 (0.25)	0.81	0.01	0.18	0.72 (0.19)	0.00	0.85	0.15	0.77 (0.18)	0.03	0.07	0.89
	20	0.995	0.95	0.86 (0.21)	0.92	0.00	0.08	0.84 (0.15)	0.00	0.95	0.05	0.87 (0.13)	0.01	0.02	0.97
	50	0.999	0.98	0.93 (0.16)	0.96	0.00	0.04	0.88 (0.12)	0.00	0.98	0.02	0.90 (0.10)	0.00	0.01	0.99
No. individuals	15	0.978	0.87	0.73 (0.24)	0.83	0.01	0.16	0.73 (0.19)	0.00	0.87	0.13	0.79 (0.17)	0.01	0.07	0.92
	60	0.991	0.93	0.82 (0.23)	0.89	0.00	0.11	0.80 (0.17)	0.00	0.93	0.07	0.85 (0.15)	0.01	0.03	0.96
No. datasets in reference table	$3x10^4$	0.990	0.92	0.83 (0.24)	0.89	0.00	0.11	0.79 (0.15)	0.00	0.95	0.05	0.87 (0.18)	0.01	0.06	0.93
	$3x10^6$	0.990	0.92	0.81 (0.23)	0.88	0.00	0.12	0.79 (0.16)	0.00	0.94	0.06	0.85 (0.17)	0.01	0.05	0.94
Statistics	Miller	0.990	0.92	0.80 (0.22)	0.89	0.00	0.11	0.79 (0.17)	0.00	0.92	0.08	0.83 (0.16)	0.01	0.05	0.94
	Beaumont	0.988	0.91	0.79 (0.23)	0.87	0.00	0.13	0.77 (0.16)	0.00	0.93	0.07	0.84 (0.17)	0.01	0.05	0.94

The effects on inferences were evaluated by calculating the mean accuracy, the mean one-versus-all AUC of the classification, the mean ($\overline{Pi}$) and the standard deviation (sd) of the posterior probability of the true scenario *i*, and the proportion of cases in which scenario *i* has the largest posterior probability (*fi*). See Materials and Methods section for details regarding default conditions. The default number of loci, individuals and datasets in the reference table are 10, 30 and $3x10^5$, respectively.

TABLE 4: Effect of erroneous prior beliefs on the ABC method

				True introduction scenario											
				Independent				Serial				Unsampled			
Type of error	Parameter distribution	AUC	Acc	$\overline{P}_{\text{indep}}$ (sd)	f_{indep}	f_{serial}	$f_{\text{unsampled}}$	$\overline{P}_{\text{serial}}$ (sd)	f_{indep}	f_{serial}	$f_{\text{unsampled}}$	$\overline{P}_{\text{ghost}}$ (sd)	f_{indep}	f_{serial}	$f_{\text{unsampled}}$
None	default	0.991	0.92	0.81 (0.22)	0.89	0.00	0.11	0.79 (0.17)	0.00	0.91	0.09	0.84 (0.15)	0.01	0.03	0.96
Demographic (test 1)	*Too large Ns values in prior*	0.942	0.78	0.69 (0.26)	0.78	0.02	0.20	0.60 (0.22)	0.03	0.70	0.27	0.74 (0.20)	0.04	0.09	0.87
	Too low Ns values in prior	0.993	0.95	0.85 (0.21)	0.92	0.00	0.08	0.85 (0.14)	0.00	0.97	0.03	0.84 (0.16)	0.00	0.05	0.95
	Too low Nf values in prior	0.974	0.78	0.93 (0.15)	0.97	0.03	0.00	0.92 (0.12)	0.01	0.98	0.01	0.39 (0.24)	0.15	0.47	0.38
	Too low DB values in prior	0.921	0.66	0.64 (0.25)	0.71	0.00	0.29	0.37 (0.22)	0.00	0.28	0.72	0.82 (0.13)	0.00	0.02	0.98
Genetic (test 1)	*Too low $\overline{\mu}$ values in prior*	0.994	0.95	0.92 (0.17)	0.95	0.00	0.05	0.83 (0.17)	0.00	0.93	0.07	0.88 (0.16)	0.00	0.04	0.96
	Too large $\overline{\mu}$ values in prior	0.952	0.82	0.67 (0.26)	0.76	0.03	0.21	0.69 (0.19)	0.01	0.85	0.14	0.71 (0.20)	0.04	0.12	0.84
	Too large P values in prior	0.988	0.92	0.80 (0.24)	0.87	0.00	0.13	0.81 (0.17)	0.00	0.94	0.06	0.83 (0.17)	0.01	0.05	0.94
	Too low P values in prior	0.99	0.92	0.81 (0.21)	0.91	0.00	0.09	0.77 (0.18)	0.00	0.90	0.10	0.83 (0.15)	0.01	0.04	0.95
	SMM	0.986	0.89	0.77 (0.23)	0.85	0.01	0.14	0.75 (0.19)	0.00	0.89	0.11	0.82 (0.16)	0.01	0.05	0.94
Historical (test 1)	*Too low Hist values in prior*	0.981	0.86	0.81 (0.23)	0.88	0.00	0.12	0.67 (0.24)	0.00	0.74	0.26	0.85 (0.16)	0.02	0.03	0.95
False source in the invaded area (test 2)	default	NR	0.86	NR	NR	NR	NR	0.72 (0.18)	0.00	0.86	0.14	NR	NR	NR	NR
False source in the native area (test 3)	default	0.949	0.75	0.41 (0.33)	0.39	0.00	0.61	0.76 (0.18)	0.00	0.90	0.10	0.85 (0.15)	0.01	0.03	0.97
Two sources (test 4)	default	NR	0.91	0.82 (0.21)	0.91	0.00	0.09	NR	NR	NR	NR	NR	NR	NR	NR
False sequence of introductions (test 5)	default	0.810	0.92	0.78 (0.23)	0.87	0.00	0.12	0.16 (0.17)	0.00	0.06	0.94	0.83 (0.16)	0.01	0.05	0.94

The effects on inferences were evaluated by calculating the mean accuracy (Acc), the mean one-versus-all AUC of the classification, the mean ($\overline{Pi}$) and the standard deviation (sd) of the posterior probability of the true scenario *i*, and the frequency (*f*i) of cases in which scenario *i* has the largest posterior probability. NR =, not relevant. SMM: Stepwise mutation model (Estoup *et al.*, 2002). See Table 1 for details regarding parameter distributions. The "prior support" is the region of the prior distribution with positive probabilities.

Titles and legends to figures

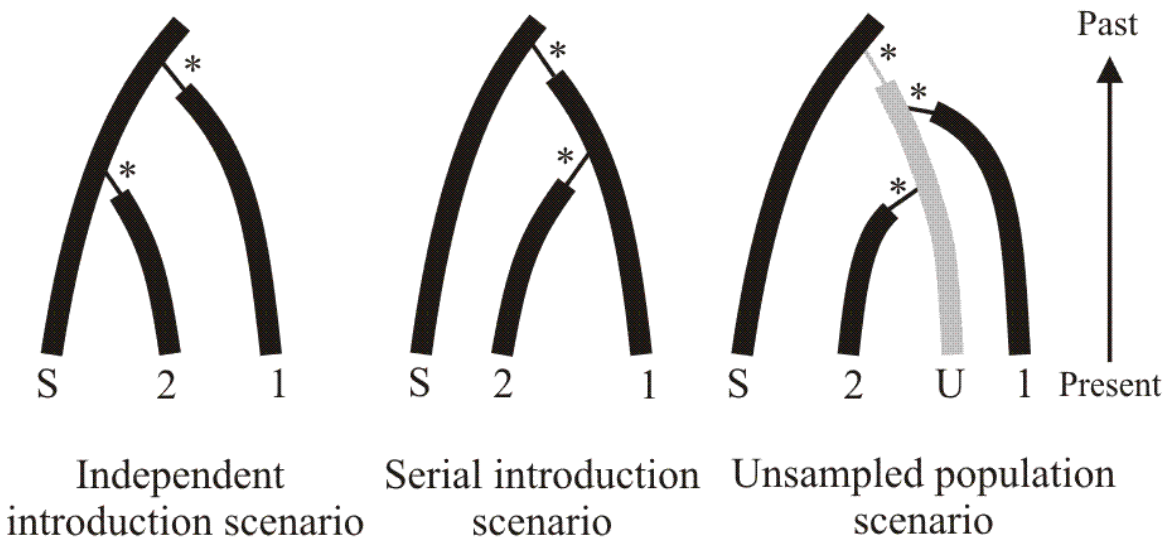
Figure 1: Introduction scenarios considered for the inference of the introduction routes of invading populations 1 and 2. S is the source population in the native area, and U, the unsampled population in the introduced area that is actually the source of populations 1 and 2 in the unsampled population scenario. The stars indicate the various bottlenecks occurring in the first few generations following introductions. The dynamics of these bottleneck events is represented in the lower part of the figure. Nf_i is the effective number of funders, BD_i, the bottleneck duration, and N_i, the effective size of invading population i. Population i was founded t_i generations before present.

Figure 2: Introduction scenarios considered in tests 2, 3, 4 and 5 of the "robustness analysis". Test 2: False source in the introduced area (the unsampled serial introduction scenario is shown as example). Test 3: False source in the native area (the false source independent introduction scenario is shown as example). Test 4: Two sources. Test 5: False sequence of introduction (the inverted serial introduction scenario is shown as example). S is the source population in the native area; U is the introduced unsampled population; FS is the false source in the native area; S1 and S2 are the two sources in the native area of the two invasive populations (in the example shown, only S2 was sampled). The grey color indicates that the corresponding population was not sampled.

Figure 3: Probability of the independent scenario in the case of the NW Italian and Central European western corn rootworm invasive outbreaks shown as a function of the F_{ST}-value between the various source populations used to compute the probability and the

Delaware sample, considered to be the “true” source of the European outbreaks (see text for details). The error bars correspond to the 95% confidence interval of the polychotomous logistic regression estimator. The black triangles correspond to the case in which three scenarios are tested (independent, serial and unsampled population), and the open squares to the case in which only two scenarios are tested (independent and serial). The serial scenario has a null posterior probability whatever the source population used and, thus, $P(\text{Unsampled}) = 1\text{-}P(\text{Independent})$.

Figure 1


Past
S
2
1
S
2
1
S
2
U
1
Present
Independent introduction scenario
Serial introduction scenario
Unsampled population scenario

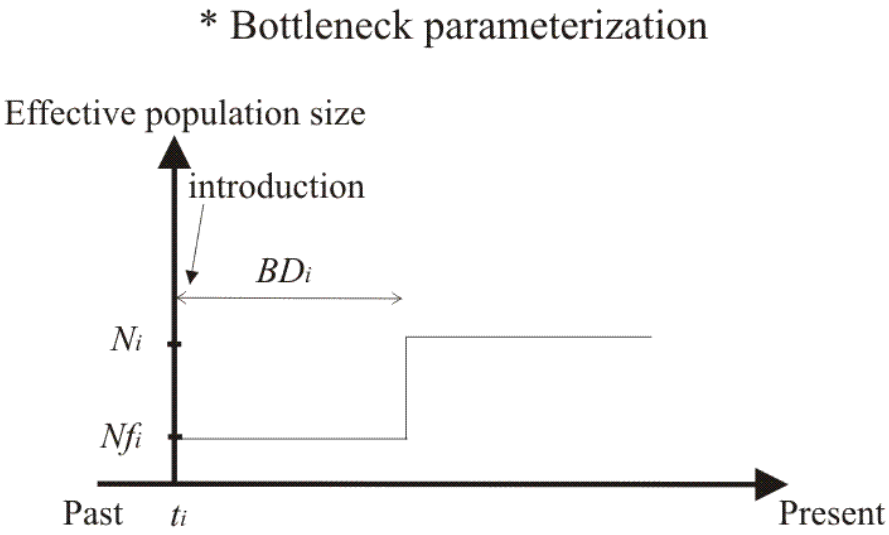

* Bottleneck parameterization
Effective population size
introduction
BDi
Ni
Nfi
Past
ti
Present

Figure 2

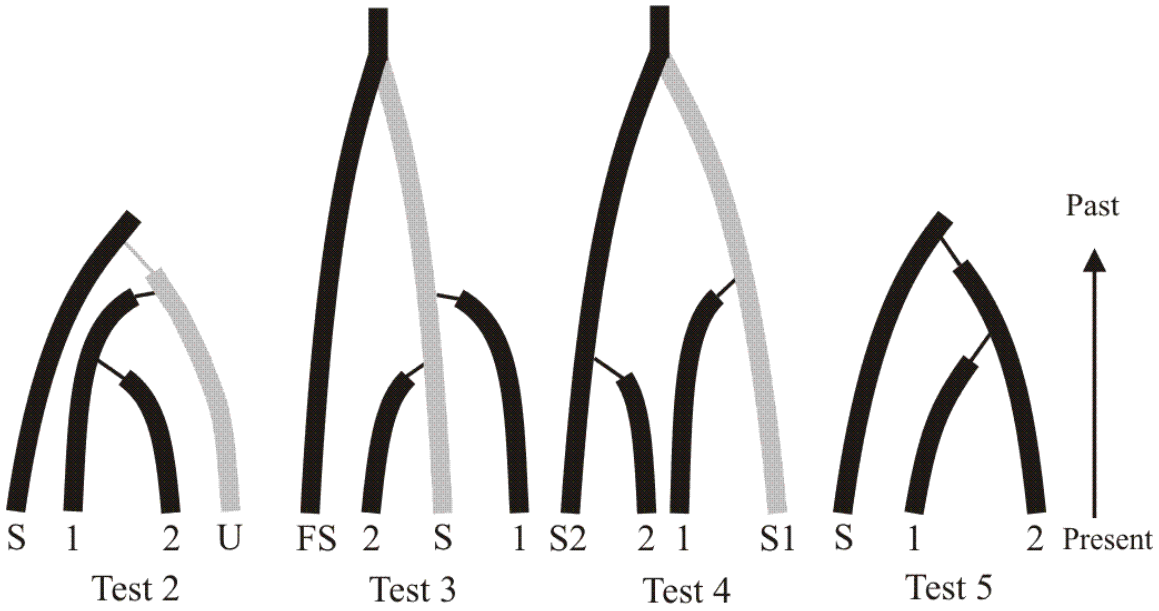
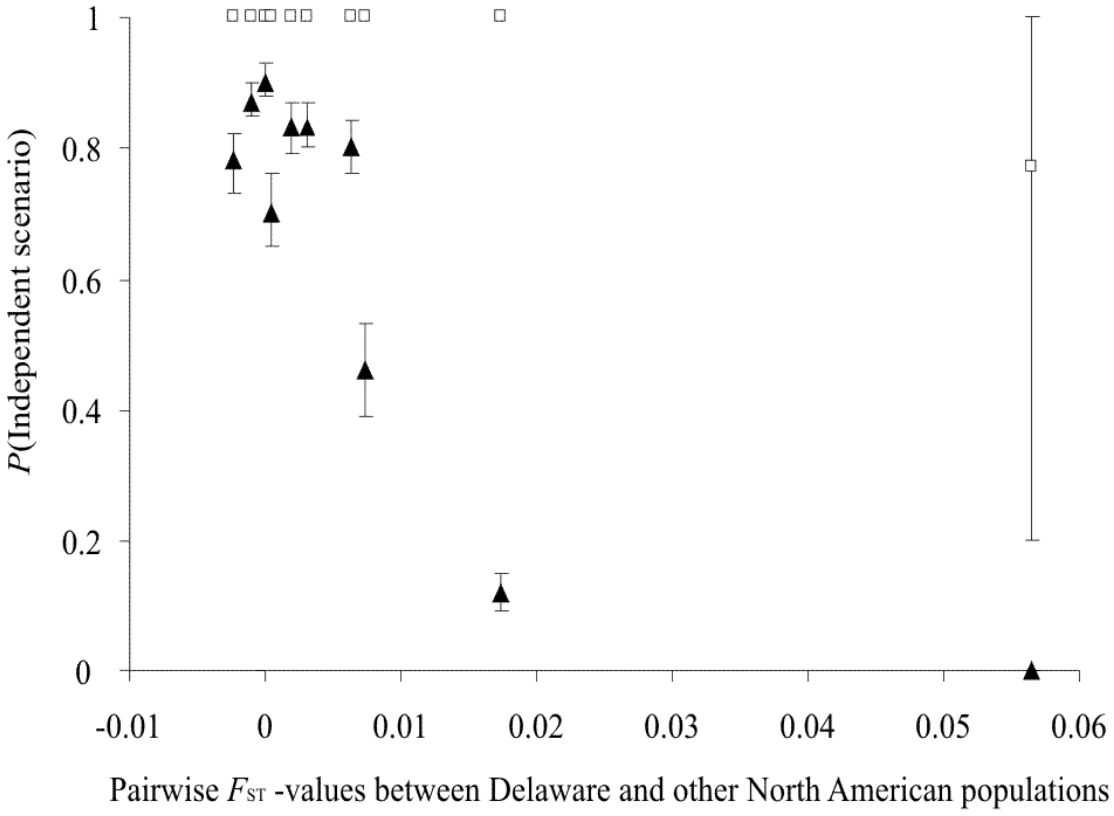


Figure 3


1
0.8
0.6
0.4
0.2
0
P(Independent scenario)
-0.01
0
0.01
0.02
0.03
0.04
0.05
0.06
Pairwise FST -values between Delaware and other North American populations

Supplementary Information

Analysis, using linear models, of the effects of scenario parameter values on posterior probabilities of scenarios

We specifically addressed the question of the effect of variation in the demographic, historical and mutational parameters used to simulate the pseudo-observed data on estimates of the posterior probabilities of the competing scenarios. We carried out a statistical analysis in which the explanatory variables were the parameters of the scenarios and the dependent variables were the posterior probability estimates. The 1,000 pseudo-observed datasets simulated per introduction scenario and the corresponding sets of parameters were used. The analysis was performed on the results obtained using the alternative parameter distributions and the alternative prior distributions (both detailed in Table 1). These alternative distributions were chosen to ensure sufficient variability of the parameters (the support of the distributions are larger than those of the default distributions), and to allow evaluation of the effect of the mutation rates and the coefficient *P* of the GSM model.

The assumption of a normal distribution of the residuals clearly did not hold true, but no improvement was obtained with other distributions (results not shown). Logit and Arcsin transformations of the dependent variable and log transformation of the explanatory variables were assessed based on the proportion of the variance accounted for by the model. We first fitted general additive models (gam function, Faraway, 2006) implemented in version 2.4.1 of R (R Development Core Team, 2006) to the data and checked for a linear relationship between the dependent variable and each explanatory variable. As this relationship was demonstrated to be linear in most cases, we then simply fitted linear models (function *lm* in R). If an explanatory variable had a non linear effect and this effect was non monotonous, we

used right and left "hockey-stick" functions of the variable (see p. 237 of Faraway, 2006) and fitted a piece-wise linear model using these functions. $\bar{\mu}$, *Ni*, and *Nfi* were log-transformed and the dependent variable was logit-transformed before the fit.

LITERATURE CITED

Faraway JJ (2006). *Extending the linear model with R*. Chapman & Hall/CRC: Boca Raton, FL.

R Development Core Team. (2006). Vienna, Austria.

Supplementary Table 1: Effect of parameter values on probability values of the true scenarios estimated with the ABC method

	True scenario											
	Independent				Serial				Ghost			
Coefficients:	Estimate	sd	*t* value	*P*	Estimate	sd	*t* value	*P*	Estimate	sd	*t* value	*P*
Intercept	-1.61	3.17	-0.51	0.61	-7.90	2.19	-3.61	**3x10^{-4}**	-3.19	4.55	-0.70	0.48
log($\overline{\mu}$)	0.59	0.14	4.22	**3x10^{-5}**	0.24	0.10	2.52	**0.01**	-0.30	0.18	-1.67	0.09
$\overline{P}$	1.19	1.24	0.96	0.34	1.77	0.85	2.09	**0.04**	0.30	1.51	0.20	0.84
log(N_{source})	0.40	0.12	3.33	**9x10^{-4}**	0.29	0.08	3.36	**8x10^{-4}**	0.25	0.15	1.67	0.09
log(N_1)	0.08	0.12	0.63	0.53	0.02	0.08	0.24	0.81	0.10	0.15	0.64	0.52
log (Nf_1)	0.89	0.10	8.82	**<10^{-6}**	1.28	0.07	17.90	**<10^{-6}**				
log(Nf_1) ≤ 3.9									2.04	0.30	-6.88	**<10^{-6}**
log(Nf_1) > 3.9									-4.82	0.25	-19.47	**<10^{-6}**
BD_1	-0.15	0.06	-2.67	**8x10^{-3}**	-0.14	0.04	-3.57	**4x10^{-4}**	0.49	0.07	6.98	**<10^{-6}**
t_1	-0.01	0.01	-1.12	0.26	4x10^{-3}	0.01	0.66	0.51	-0.01	0.01	-0.50	0.62
log(N_2)	-0.16	0.12	-1.34	0.18	0.11	0.08	1.36	0.17	0.04	0.15	0.25	0.80
log(Nf_2)	1.25	0.11	11.87	**<10^{-6}**	0.68	0.07	9.36	**<10^{-6}**	1.57	0.13	11.90	**<10^{-6}**
BD_2	-0.12	0.06	-2.12	**0.03**	-0.05	0.04	-1.33	0.18	-0.17	0.07	-2.37	**0.02**
t_2	-0.01	0.01	-0.71	0.48	4x10^{-3}	0.01	0.04	0.96	0.01	0.01	1.37	0.17
log($N_{unsampled}$)									-0.06	0.15	-0.40	0.69
log($Nf_{unsampled}$) ≤ 4.6									3.26	0.20	-16.01	**<10^{-6}**
log($Nf_{unsampled}$) > 4.6									-3.51	0.52	-6.69	**<10^{-6}**
$BD_{unsampled}$									0.02	0.07	0.21	0.83
$t_{unsampled}$									-4x10^{-3}	0.02	-0.27	0.79

N_{source} is the effective population size of the source, Nf_i, the effective number of funders, BD_i, the bottleneck duration, N_i, the effective size after growth of outbreak i (with i = 1, 2 or unsampled). Outbreak i was founded t_i generations before present. $\overline{\mu}$ is the mean mutation rates and $\overline{P}$ is the mean coefficient of the geometric distribution of repeat units. Effects on the logit-transformed posterior probabilities of the true scenarios were assessed with linear models. *P* is the *p*-value associated with the estimate of the parameter coefficient. Indices of parameters refer to the population numbers and names used in Figure 1. Probabilities below 0.05 are shown in bold character.

Supplementary Table 2: Performance of the ABC method when four competing introduction scenarios are considered

True scenario	Tested scenario	f	f_{lim}	$\overline{P}$ (sd)
Independent	Independent	0.88	0.85	0.73 (0.26)
	Serial	0.01	0.01	0.06 (0.08)
	Inverse serial	0.02	0.01	0.07 (0.10)
	Unsampled	0.09	0.06	0.14 (0.14)
Serial	Independent	0.00	0.00	0.00 (0.02)
	Serial	0.91	0.85	0.68 (0.18)
	Inverse serial	0.02	0.02	0.09 (0.09)
	Unsampled	0.07	0.03	0.23 (0.12)
Inverse serial	Independent	0.00	0.00	0.00 (0.02)
	Serial	0.02	0.01	0.10 (0.09)
	Inverse serial	0.91	0.85	0.67 (0.18)
	Unsampled	0.07	0.04	0.23 (0.12)
Unsampled	Independent	0.01	0.01	0.02 (0.07)
	Serial	0.05	0.04	0.22 (0.11)
	Inverse serial	0.07	0.03	0.22 (0.11)
	Unsampled	0.87	0.79	0.55 (0.11)

f is the proportion of the simulated data classified into each tested scenario. The largest posterior probability defines the chosen scenario. The frequency at which the lower limit of the 95% confidence interval of the posterior probability of each scenario exceeds the upper limits of the other scenarios is f_{lim}. The mean ($\overline{P}$) and the standard deviation (sd) of the posterior probability of each tested scenario are shown.

Supplementary Figure 1: Effect of an error concerning the source in the area of origin

Pseudo-observed datasets (n = 1,000) were simulated under the independent introduction scenario. Box plots of posterior probability values for the independent (white bars) and unsampled (gray bars) introduction scenarios are presented as functions of the F_{ST} ranges measured between the true and false source population. Posterior probabilities for the serial introduction scenarios are all close to zero and are therefore not presented.