

ORIGINAL ARTICLE

Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source

A Benazzo¹, S Ghirotto¹, ST Vilaça¹ and S Hoban^{1,2}

The introduction of invasive species to new locations (that is, biological invasions) can have major impact on biodiversity, agriculture and public health. As such, determining the routes and modality of introductions with genetic data has become a fundamental goal in molecular ecology. To assist with this goal, new statistical methods and frameworks have been developed, such as approximate Bayesian computation (ABC) for inferring invasion history. Here, we present a model of invasion accounting for multiple introductions from a single source (MISS), a heretofore largely unexplored model. We simulate microsatellite data to evaluate the power of ABC to distinguish between single and multiple introductions from the same source, under a range of demographic parameters. We also apply ABC to microsatellite data from three invasions of bumblebee in New Zealand. In addition, we assess the performance of several methods of summary statistics selection. Our simulated results suggested good ability to distinguish between one- and two-wave models over much but not all of the parameter space tested, independent of summary statistics used. Globally, parameter estimation was good except for bottleneck timing. For one of the bumblebee species, we clearly rejected the MISS model, while for the other two we found inconclusive results. Since a second wave may provide genetic reinforcement to initial colonists, help relieve inbreeding among founders, or increase the hazard of the invasion, its detection may be crucial for managing invasions; we suggest that the MISS model could be considered as a potential model in future theoretical and empirical studies of invasions.

Heredity (2015) **115**, 262–272; doi:10.1038/hdy.2015.38; published online 29 April 2015

INTRODUCTION

The introduction of species to new locations can have substantial ecological and evolutionary consequences for both the non-native species and the receiving ecological community. Furthermore, some introduced species became invasive, a major environmental challenge threatening natural ecosystems (Lankau *et al.*, 2009), the services provided by ecosystems to humans such as agricultural production (Boubou *et al.*, 2013) and human health (Fonseca *et al.*, 2010). As climate changes and globalization increases in the future, introductions and invasions will also increase (Diez *et al.*, 2012).

In the past decade, genetic data have increasingly been used to successfully describe invasion processes, especially population dynamics early in an introduction, including the source and number of introduced individuals (Clegg *et al.*, 2002; Ficetola *et al.*, 2008), the time during which introduced populations remain small (bottleneck duration or lag time, Lye *et al.*, 2011) and adaptive evolutionary changes (Lee 2002; Phillips *et al.*, 2006). Elucidating the details of these invasion processes can aid in controlling particular species, such as choosing biological or chemical controls specific to the source or designing monitoring and inspection programs for particular trade routes or expected propagule size (Lodge *et al.*, 2006; Estoup and Guillemaud, 2010). As more invasions are described, a clearer picture is developing regarding characteristics of successful and unsuccessful invasions, including typical propagule sizes, number of introductions, admixture rates and rate of spread (Wilson *et al.*, 2008), though the

demographic, ecological and evolutionary factors contributing to invasions are still unresolved and are a major area of research.

Accumulated genetic evidence not only suggests that invasions often feature complex histories, but also highlights several key and relatively simple components (Guillemaud *et al.*, 2011). For example, a common finding is that multiple invading populations may originate from multiple geographically distinct sources (Miller *et al.*, 2005; Keller, 2009; Guillemaud *et al.*, 2010). An alternative and also common scenario for the origin of multiple invading populations is a spatial expansion/serial founding from a single source, for example, an invasion front (Clegg *et al.*, 2002; Estoup *et al.*, 2004) or a bridgehead process (Guillemaud *et al.*, 2011; Lombaert *et al.*, 2012), in which one successfully established population is the source for many other populations. Lastly, admixture between distinct source populations is common between invaded populations that have existed for some time due to migration exchange or merging (Besnard *et al.*, 2014).

An additional possible scenario, which is rarely considered is a model that we term ‘multiple introductions from the same source’ (hereafter called MISS). This scenario concerns one invasive population. An introduced species establishes at a single site. After a small number of generations, a second wave of introduced individuals arrives to the site. The novel aspect of this model is that the second wave originates from the same source as the first; most models of multiple introductions consider a second wave originating from a

¹Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy and ²National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, Knoxville, TN, USA
Correspondence: Dr S Hoban, National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, 1122 Volunteer Blvd, Knoxville, TN 37996, USA.
E-mail: shoban@alumni.nd.edu

Received 12 November 2014; revised 12 March 2015; accepted 14 March 2015; published online 29 April 2015

geographically and genetically different source than the first (Miller *et al.*, 2005; Estoup and Guillemaud, 2010). (The MISS model could also include multiple introductions from separate geographic sources if the introduced species is entirely genetically homogenous in its native range.) The MISS scenario is plausible and perhaps likely; if an introduction via a particular pathway or vector happens once (deliberately or accidentally) and the pathway or vector (movement of cargo, escape of a cultivated species) persists, a second introduction may occur (Lodge *et al.*, 2006).

For example, bumblebees were introduced from the United Kingdom to New Zealand in two waves, in 1885 and 1906. In this case, Lye *et al.*, 2011 asserted that this MISS scenario may be difficult to detect: ‘preliminary power analysis based on simulated data sets indicated that it would not be possible to produce accurate parameter estimates from a two-step introduction scenario (results not shown)’. Guillemaud *et al.*, 2010 also recognized the second wave as a possibility but their model nonetheless ‘assume(s) that no repeated introductions occurred at the same location’. In addition, the ‘late second wave’ was considered in human colonization of the Americas (Ray and colleagues, 2010). A secondary introduction has also been discussed with respect to crop domestication (Olsen and Gross, 2008), and may have occurred during recent cactus invasions (Marsico *et al.*, 2010). Biocontrol agents, escaped pets or pests may be especially likely to establish in two or more distinct waves over time. To our knowledge, while the MISS model has been acknowledged as a possibility, it has not been widely explored in models of invasive species.

Our goal in this paper was to evaluate the ability of genetic data jointly with the approximate Bayesian computation (ABC, explained in Methods section) framework to detect a second wave from the same source. To achieve this goal, we simulated genetic datasets of an invasion produced by a single wave (null model) and two waves (alternative model) of colonization, across a wide range of parameters (time of introduction, length of bottleneck and number of colonizer individuals). Simulated data have been used successfully to test other invasion models (for example, ghost populations) and the power of different analyses and sampling strategies (Pascual *et al.*, 2007; Muirhead *et al.*, 2008; Guillemaud *et al.*, 2010). We tested whether ABC analysis can establish statistical support for the second-wave scenario by calculating power and type I error. We expected that some second waves may be more easily detectable than others, so our first aim was to roughly define the parameter space in which the second-wave model can be distinguished. We then quantified how well parameters can be estimated, and we tested moderate and large numbers of loci. We also compared methods for choosing summary statistics. Lastly, we applied the ABC method to a microsatellite dataset with three invading bumblebee species in New Zealand.

MATERIALS AND METHODS

Approximate Bayesian computation

We used ABC to obtain the posterior probability of the demographic models under investigation and the posterior distribution of each parameter characterizing the models given the genetic data (Beaumont *et al.*, 2002; Bertorelle *et al.*, 2010). ABC can be applied to study models for which the likelihood function is not available or too hard to compute. Under ABC, millions of genetic datasets with particular features (for example, number of individuals and number/type of genetic markers) are generated according to specific demographic models by drawing model parameters from the associated prior distributions. The pattern of genetic variation in the observed (or pseudo-observed, for our study) and simulated data, captured by a number of summary statistics, is then compared. Only the demographic parameters (that is, simulations) that generated summary statistics closest to the observed ones are considered to calculate parameters posterior distribution and model posterior probabilities (for

extensive explanation of ABC, see Beaumont, 2010; Bertorelle *et al.*, 2010). We used ABCsampler in the software ABCtoolbox (Wegmann *et al.*, 2010) for generating reference tables and custom scripts for parameter and models probability estimation.

The model

Our model is shown in Figure 1, with parameters listed in Table 1. Briefly, at some time in the recent past (TS), a number of colonizers (small or moderate number, NB) is introduced from a single source population, and experiences a time (of short or moderate length) at small and constant population size (a bottleneck, TB). Following this time period, a second wave of colonizers arrives (TM) and mixes with the first. The number of individuals of the two waves is assumed to be equal; while this is unlikely to be exactly true, in many cases, if the waves arrive by the same pathway (for example, ship's ballast), the size of the waves might be similar. Thereafter, the invasive population grows exponentially to the present time (NI, some multiple of NB). The time point at which exponential growth begins is referred to as TG, and the population growth rate is GROWTH. A bottleneck period and subsequent exponential growth are common features of invasion models (Estoup and Guillemaud, 2010; Guillemaud *et al.*, 2011). Meanwhile the source population remains at constant size (NS). All population sizes refer to effective sizes. The null model includes the same series of events (introduction, bottleneck and growth) but no second wave. We assumed that the correct source is sampled and that no structure is present in the source population, there is no connection via migration to other invading populations, and equal sample sizes are taken from the invading and source population. Implications of not sampling the true source and of sample size are explored by other authors (Muirhead *et al.*, 2008; Guillemaud *et al.*, 2010). Simulations were made with simcoal2 (Excoffier and Foll, 2011), which explicitly allows multiple coalescent events when population sizes are small. Our simulated data are similar to a range of empirical datasets (Supplementary Document 1).

Prior distributions

We concentrated on a widely used marker in conservation genetics (that is, microsatellites/short tandem repeats), simulating all scenarios with either 20 or 100 unlinked loci. Although 100 loci are many, genomic resources for many species are rapidly expanding, particularly for species of high economic concern (for example, invaders such as rat, Argentine ant, honey bee, rainbow trout and loblolly pine), so this number is reasonable. Locus-specific mutation rates for short tandem repeats were drawn from a gamma distribution (mean = 0.0005, shape parameter = 2) to cover a range of mutation rates (Excoffier *et al.*, 2005; Neuenschwander *et al.*, 2008). The mean mutation rate we refer to as MU.

First, we used coalescent simulations to create 1000 pseudo-observed datasets (PODs) under each of eight possible MISS scenarios (Table 2; Figure 2)—all combinations of moderate and small introduction, long and short bottleneck, and older and more recent introductions. Considering the wide parameter space of these eight scenarios should help determine how robust the ABC

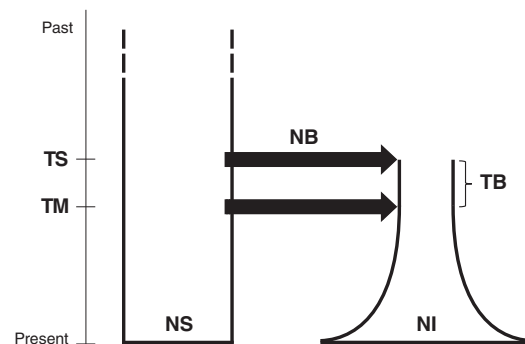


Figure 1 Schematic of MISS. NB, population size of wave; NI, current population size of invading population; NS, current population size (*N*) of source; TB, length of time of bottleneck; TM, time of wave of invasion (migration); TS, time of second wave.

Table 1 Parameters in the model and prior parameter values for ABC reference table

Parameter	Abbreviation	Prior distribution
Population size of the source	NS	Uniform (50 000:200 000)
Current size of invading population	NI	Uniform (5000:400 000)
Number of individuals introduced (bottleneck)	NB	Uniform (4:400)
Time (before present) of first introduction (time of split)	TS ^a	Uniform (15:400)
Length of bottleneck phase	TB	Uniform (1:80)
Time (before present) of end of bottleneck, beginning of exponential growth	TG	TS – TB
Time (before present) of second wave (migration event), if existing	TM	TS – TB – 1
Mean mutation rate	MU ^b	Uniform (0.001:0.0001)
Growth rate	GROWTH	1/(TS – TB) ln(NB/NI)

Abbreviation: ABC, approximate Bayesian computation.

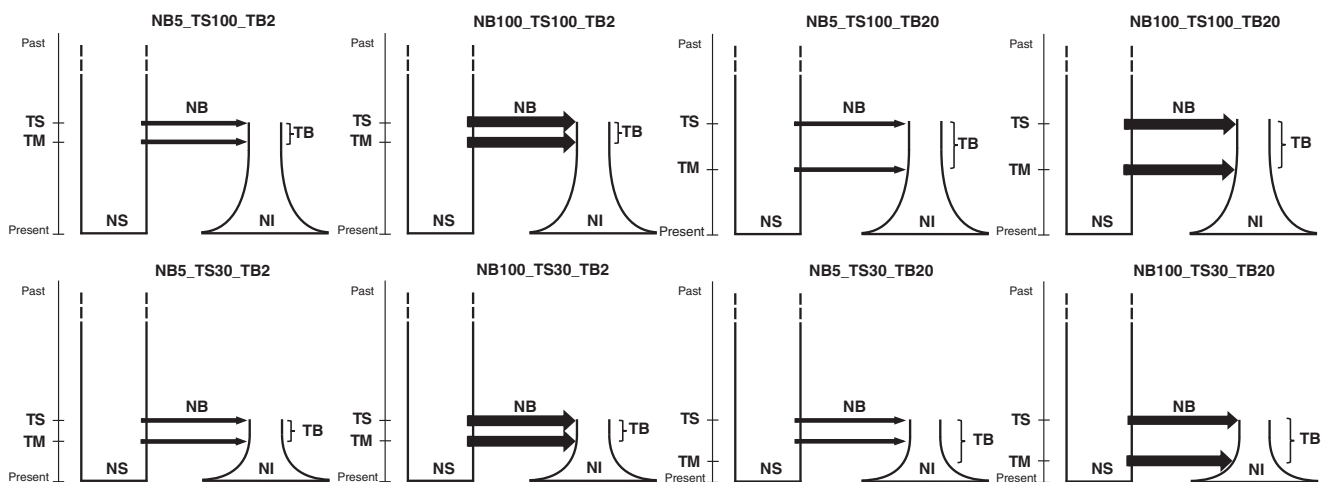
^aSubject to constraint: TS > (TB + 1).

^bThe mean for the gamma distribution (see text).

Table 2 The eight MISS scenarios considered in our study

Number	Description	TS	TB	NB	Abbreviation
1	Older introduction, short bottleneck and small founding size	100	2	10	NB10_TS100_TB2
2	Older introduction, short bottleneck and moderate founding size	100	2	200	NB200_TS100_TB2
3	Older introduction, long bottleneck and small founding size	100	20	10	NB10_TS100_TB20
4	Older introduction, long bottleneck and moderate founding size	100	20	200	NB200_TS100_TB20
5	Recent introduction, short bottleneck and small founding size	30	2	10	NB10_TS30_TB2
6	Recent introduction, short bottleneck and moderate founding size	30	2	200	NB200_TS30_TB2
7	Recent introduction, long bottleneck and small founding size	30	20	10	NB10_TS30_TB20
8	Recent introduction, long bottleneck and moderate founding size	30	20	200	NB200_TS30_TB20

Abbreviations: NB, number introduced; TB, length of bottleneck phase; TS, time of introduction.

**Figure 2** Visual of eight scenarios considered.

procedure can be expected to perform in real datasets. The parameter values chosen for simulations are similar to parameters of real invasions of mammals, insects, amphibians and plants (Supplementary Document 2). In these PODs, parameter values are fixed. Nonetheless each simulated dataset was distinct due to mutational, genealogical and sampling stochasticity.

Second, we analyzed each of the 16 000 PODs (1000 for each of the eight scenarios, for two models) with an ABC procedure to quantify statistical support for one of two models (single introduction or second wave), and to estimate the associated demographic parameters. Thus for each scenario our results include power for model choice and measures of performance (for example, bias) for parameter estimation.

To analyze each of the PODs created under our scenarios mentioned above ('local' scenarios), the ABC procedure requires a reference table covering plausible values (priors) for the invasion. We simulated one million datasets for

the null model reference table and one million datasets for the second-wave reference table. Priors (Table 1) were chosen to widely encompass all scenarios considered, with the lower bound as one-half the minimum value considered in the eight scenarios and the upper bound as twice the maximum value considered. We used uniform priors because the plausible parameters covered at most two orders of magnitude. The reference table will also be used to estimate performance 'globally', for example, across the entire range of plausible parameters, see Performance section.

Summary statistics

We compiled a large number of summary statistics (Supplementary Document 3), which, a priori, we expected may be influenced by the bottleneck, migration and growth scenario. We used Arlequin 3.5 (Excoffier and Lischer, 2010), in-house code and R packages mmod (Winter 2012) and pegas (Paradis 2010)

to calculate statistics. We chose the following ‘within-population’ and ‘between-population’ statistics, sometimes called one-sample (or diversity) and two-sample (or differentiation) statistics: number of alleles, allelic range, expected heterozygosity, the Garza and Williamson’s M-ratio (Garza and Williamson, 2001), number of private alleles, number of shared alleles, Fis, assignment likelihood ratio (Paetkau *et al.*, 1995), Rst, delta mu squared, Jost’s D, Nei’s GST, Hedrick’s GST and Weir and Cockerham’s FST and FIT. We calculated mean and s.d. of each. From this large set, we then chose a small number that were found to be common in a review of recent ABC studies (Bertorelle *et al.*, 2010). Hereafter, we refer to these as the ‘full set’ and ‘minimal set’ of summary statistics (Supplementary Document 3). We also created a ‘ranked’ set of 5 and 10 statistics, as we describe in the next paragraph. We performed power calculation (see below, with a range of decision thresholds) with each of the four sets.

Choosing an appropriate set of summary statistics is a current area of interest in ABC methodology (Csilléry *et al.*, 2010). As argued in Veeramah *et al.*, 2012, powerful statistics may be the statistics having the least overlap in their distributions under the two opposing models. We therefore performed a Kolmogorov-Smirnov (a Kruskal-Wallis test on distribution means was also performed yielding highly similar results) test to compare these distributions for each statistic (Supplementary Figure 1) and obtain a *P*-value (as in Veeramah *et al.*, 2012). We ranked statistics by *P*-value and chose the 5 and 10 most significant.

Model choice/power calculation

For each POD, we determined the posterior probability for the two opposing models (null and second wave) through a polychotomous logistic regression (Beaumont, 2008) for the 20 000 best simulations. We determined power to choose the correct model for each of the eight scenarios, created under both the null and the second-wave model, as the proportion of PODs that were assigned to the correct model (that is, rate of true positives). To assign each POD to a model, we defined a set of probability thresholds, that is, the threshold at which one model is chosen as the true one. When this threshold was 0.5, the model with the higher probability was considered the supported model even though the difference between the two posterior probabilities can be extremely low (that is, 0.51 vs 0.49). When higher thresholds were considered (for example, 0.7; Table 3), power was computed as the proportion of PODs that exhibited, in one of the two compared models, posterior probabilities higher than the threshold. We also calculated power with the full range of possible decision probability thresholds (0.5 and 0.7 shown in Table 3; from 0.55 to 0.95 in Supplementary Document 4). Higher decision thresholds are more statistically conservative, and thus higher thresholds should result in a decrease in power.

Further simulations

We identified high error rates in model choice for two null and two second-wave scenarios (see Results section). Therefore, we performed further analyses to explore these scenarios in detail (explained more in Supplementary Methods). First, we performed simulations with Bayesian Serial SimCoal (Anderson *et al.*, 2005) in order to inspect summary statistics variation through time, and to determine if high drift might erase genetic signals of a second wave. Second, we performed additional simulations with TB=5 and TB=10 generations to see if moderate bottleneck times show performance in between the 2-generation and 20-generation bottlenecks (our main simulation scenarios).

Parameter estimation and summary statistics

As in model choice, during parameter estimation we searched for a best set of summary statistics. We tested two alternative methods for choosing a best set of summary statistics and one method for data reduction; these are just three examples of methods developed to improve parameters estimation in ABC (see also Aeschbacher *et al.*, 2012 and Stocks *et al.*, 2014, described in Supplementary Methods). (1) Using all datasets in the reference table, we calculated the coefficient of determination (*R*²) of a linear regression of each summary statistic on each model parameter (Hamilton *et al.*, 2005). We then ranked statistics by their *R*² values and used statistics with the highest values for parameter estimation. (2) We used the minimum entropy selection criterion, implemented in the R package ABCme (Nunes and Balding, 2010). Under this framework, for each parameter, all factorial combinations of a given set of statistics are tested iteratively to estimate the set that minimizes the entropy (that is, the square root of the sum of squared errors) of the resulting posterior distribution (for details see Supplementary Methods). (3) We transformed all summary statistics into a vector of partial least square (PLS) components (Tenenhaus *et al.*, 1995; Mevik and Wehrens, 2007). The PLS method transforms summary statistics (that is, predictor variables) in a reduced set of orthogonal components maximizing the covariance matrix of predictor and response variables, to better explain the variance of the parameters. This transformation may avoid the ‘curse of dimensionality’, arising when too many summary statistics are used to under the ABC framework. To determine the appropriate number of PLS components for each parameter in our model, we visually examined the graphs of root mean squared error (RMSE) against the number of components, as suggested by Wegmann *et al.*, 2010.

Performance

The summary statistics/PLSs from these three methods were used to estimate parameters and compare estimates to known values using various measures of performance. To measure performance, we used relative bias, RMSE, coverage,

Table 3 Probability support for NULL or SW model for each of the eight scenarios considered using all the available statistics for moderate and large NL

Threshold	NL	NB10_TS30_TB2		NB10_TS30_TB20		NB10_TS100_TB2		NB10_TS100_TB20		NB200_TS30_TB2		NB200_TS30_TB20		NB200_TS100_TB2		NB200_TS100_TB20	
		NULL	SW	NULL	SW	NULL	SW	NULL	SW	NULL	SW	NULL	SW	NULL	SW	NULL	SW
<i>(A) Simulated data created under the NULL model</i>																	
0.5	20	0.998	0.002	1.000	0.000	1.000	0.000	1.000	0.000	0.003	<u>0.997</u>	0.953	0.047	0.116	<u>0.884</u>	0.796	0.204
0.7	20	0.968	0.000	1.000	0.000	0.996	0.000	1.000	0.000	0.000	<u>0.933</u>	0.676	0.003	0.022	<u>0.667</u>	0.342	0.022
0.5	100	0.994	0.006	1.000	0.000	1.000	0.000	1.000	0.000	0.211	<u>0.789</u>	1.000	0.000	0.098	<u>0.902</u>	0.982	0.018
0.7	100	0.931	0.000	1.000	0.000	0.998	0.000	0.999	0.000	0.064	<u>0.517</u>	0.995	0.000	0.022	<u>0.692</u>	0.793	0.000
<i>(B) Simulated data created under the SW model</i>																	
0.5	20	0.473	0.527	0.000	1.000	<u>0.997</u>	0.003	<u>0.981</u>	0.019	0.000	1.000	0.000	1.000	0.047	0.953	0.036	0.964
0.7	20	0.303	0.332	0.000	1.000	<u>0.992</u>	0.003	<u>0.972</u>	0.013	0.000	0.988	0.000	0.982	0.005	0.785	0.002	0.875
0.5	100	0.193	0.807	0.000	1.000	<u>0.996</u>	0.004	<u>0.982</u>	0.018	0.007	0.993	0.000	1.000	0.063	0.937	0.022	0.978
0.7	100	0.047	0.537	0.000	1.000	<u>0.987</u>	0.001	<u>0.949</u>	0.012	0.000	0.960	0.000	0.996	0.015	0.818	0.003	0.939

Abbreviations: NL, number of loci; NULL, null; SW, second-wave model. The supported model is shown in italics. When the supported model is not the correct one, values are underlined. For example in A for NB10_TS30_TB2, the NULL model is supported (correctly), while for NB200_TS30_TB2 the SW model is (incorrectly) supported.

range and factor 2. Measures of performance were calculated globally (estimating parameters for 1000 randomly chosen simulated datasets for each model) and locally (estimating parameters for the 1000 PODs created under each of our eight scenarios, for each model). Global performance reflects how well the second wave can be described across the wide range of parameters considered (the entire range of priors), while local performances show how well the second wave can be described for each specific scenario (for example, short bottleneck and large second wave; Table 1).

Application to bumblebee introduction in New Zealand

Bumblebees were introduced into New Zealand, for pollination purposes, in two documented events in 1885 and in 1906 from Britain. No information are available about the exact composition of the introduced individuals but it is believed that they belonged to at least six species. Four species became established (*Bombus terrestris*, *Bombus hortorum*, *Bombus ruderatus* and *Bombus subterraneus*) increasing their number from few hundreds to thousands of individuals. Three species (*B. terrestris*, *B. hortorum* and *B. ruderatus*) are currently present in Britain and can be considered as the 'source' populations. By using this case study, Lye *et al.*, 2011 studied the demographic history of the three species in New Zealand using nuclear microsatellite data assuming a single introduction event for each. The authors state that a power analysis indicated no power to correctly estimate the parameters of a two-step introduction scenario using ABC.

We explicitly quantified support for single vs multiple introduction models (MISS), under the ABC framework, for the three bumblebee species, using the eight microsatellite from Lye *et al.*, 2011 available in Dryad (doi:10.5061/dryad.sk22v). For each species, we generated one million simulated datasets for the single and multiple introduction models using sample sizes described in Table 5, a generation time of 1 year (Lye *et al.*, 2011) and using the default set of priors present in Table 2. The polychotomous logistic regression approach based on the best 20 000 simulations was used to obtain an estimate of the posterior probability for the two tested models.

RESULTS

Model choice

Use of 20 loci and 100 loci considering all summary statistics showed highly similar results (Table 3; Supplementary Document 5), though with slightly higher power at 100 loci for stricter thresholds for some scenarios (Supplementary Document 4). We found that, of the eight possible scenarios, two scenarios were problematic for identifying the second wave and two were problematic for the null scenario (explored in detail below); the rest allowed very good ability to choose the correct model, for data simulated under both the null and alternative model (those not allowing good choice are underlined in Table 3; see also Supplementary Documents 4 and 5). For these situations, for the second-wave datasets, power at the 0.5 threshold was always >0.80 , and for the null datasets, power was always >0.95 using the full set of statistics. High power was still present using 20 loci (>0.90) except for the NB10_TS30_TB2 (abbreviations in Table 1), where we detected a slight decrease in power until 0.53. At a stricter threshold (0.80; Supplementary Document 4), four second-wave scenarios still performed very well (power always >0.85), as do four null scenarios (power ≥ 0.95). Several even performed well at 0.90 or 0.95 thresholds. The NB10_TS30_TB20 scenario was the easiest scenario for detecting a second wave, as it was identified correctly even at the strictest threshold, for both models. Globally, long bottleneck duration (TB20) facilitated the detection of the null (single wave) model with a power always higher than 0.98 and 0.79, considering a threshold of 0.5, for both 100 and 20 loci. However, models with large bottleneck size (NB200) were always well selected, with a power >0.93 for a threshold of 0.5, irrespective to the number of loci that were considered.

As expected, when the bottleneck period was small (TB=2), it was very difficult to support the true model; in these cases, indeed, the ABC procedure identified with high confidence the other (incorrect) scenario as having generated the simulated data. In one situation (NB10_TS100_TB2), the analysis of PODs produced under the second-wave model supports with high confidence (power >0.99) the NULL model, indicating that the drift period was too short to leave a detectable signal even with 100 loci. An identifiability issue also appeared for two parameter combinations (NB200_TS30_TB2 and NB200_TS100_TB2) where the second-wave model was highly supported when the PODs came from the null scenario.

For error-prone situations (that is, cases in which the wrong model was supported by the ABC procedure), a stricter threshold helped somewhat, especially for simulations under the null model: with a 0.8 threshold, the probability to erroneously support the second-wave model was only 0.331 and 0.514, while for five of the eight scenarios simulated under the null model power remains >0.75 . Unfortunately this threshold did not help for datasets under the second-wave scenario (Table 3; Supplementary Documents 4 and 5). Following results will mainly concern 100 loci.

Interestingly, though we used both intuition and established methods to choose smaller sets of summary statistics, our minimal, five-statistic and 10-statistic sets showed little difference in power with the all-statistic set, with difference in power rarely exceeding 0.05 across the range of thresholds (Supplementary Document 4). The five-statistic set provided slightly higher power for identifying second-wave scenarios while simultaneously slightly decreasing power for the null scenarios. Specifically, for the NB10_TS30_TB2 and NB200_TS100_TB2 situations, power was much higher (0.92 and 0.996) at the 0.80 threshold with the five-statistic set than any other set. However, power to detect the null scenario correctly was substantially reduced for the NB200_TS100_TB20 scenario, from 0.424 to 0.009, and very slightly reduced under other scenarios. Importantly, for the error-prone situations, the smaller sets of summary statistics actually increased the error rate, making it more likely the incorrect model would be supported.

Further simulations

For several summary statistics, we observe that high genetic drift in two scenarios (few founders and long bottleneck) may reduce genetic diversity so severely as to erase any possible signal of two waves. Summary statistics under the null and second-wave simulations showed distinct distributions over time under the NB10_TS30 (allowing to distinguish between the null and second-wave model) but not the NB10_TS100 situation (Supplementary Methods).

Simulations of moderate bottlenecks (TB=5, 10) also showed predicted results (increased power over TB=2). With simulated data created under the null model of NB200_TS30_TB5, NB200_TS30_TB10 and NB200_TS100_TB10, the null model was correctly chosen, with power of 0.77, 0.974 and 0.802, respectively (at a decision threshold of 0.5). However, the second wave was incorrectly chosen for NB200_TS100_TB5. It is apparent that as TB increases, for these large introductions, the ability to distinguish between models increases (Supplementary Document 4).

Parameter estimation and summary statistics

Overall, the regression, PLSs and minimum entropy approaches to choosing summary statistics were in agreement (Table 4). A number of parameters were precisely and correctly inferred for all the three approaches, while other parameters were poorly predicted.

Table 4 Measure of performance for parameter estimation. Estimations made with the best method based on lowest RMSE, which was R^2 for second wave and minimum entropy/MBFV for the null model

	<i>GROWTH</i>	<i>MU</i>	<i>NB</i>	<i>NI</i>	<i>NS</i>	<i>TB</i>	<i>TG</i>	<i>TM</i>	<i>TS</i>
<i>Second wave</i>									
Bias	-0.094	-0.157	0.177	0.368	-0.088	0.687	0.164	0.214	0.133
RMSE	0.360	0.296	0.948	1.957	0.276	3.144	0.996	1.600	0.456
Factor 2	0.911	0.902	0.889	0.858	0.893	0.857	0.889	0.884	0.876
Coverage	0.812	0.894	0.941	0.688	0.939	0.701	0.898	0.893	0.900
<i>Null model</i>									
Bias	-0.004	-0.131	0.240	-0.205	-0.164	0.648	-0.043	—	-0.018
RMSE	0.601	0.293	0.634	1.500	0.317	2.204	0.645	—	0.375
Factor 2	0.647	0.900	0.936	0.318	0.837	0.793	0.769	—	0.930
Coverage	0.929	0.896	0.899	0.867	0.886	0.886	0.852	—	0.869

Abbreviation: RMSE, root mean squared error.

Regression: For the second-wave model, MU (see Table 1 for all abbreviations), NB, NS, TS, TG and TM were well predicted (high R^2), while NI, GROWTH and TB showed increasingly poor prediction, with essentially no ability for the latter two (Supplementary Documents 6 and 7). Depending on the parameter, between 5 and 15 summary statistics had coefficients of determination >0.20 . Each parameter had its own ‘best’ set of statistics, but the following commonly had high R^2 : assignment likelihood ratio for population 1 and 2, Garza and Williamson’s M-ratio of population 2, GST, FST and number of alleles shared between populations. For the null model, MU, NB, NS and TB were well predicted, while TS, TG, NI and GROWTH showed increasingly poor prediction, with essentially no ability for the latter two. Similar to the second-wave model, depending on the parameter, between 5 and 20 summary statistics had $R^2 >0.20$.

Minimum entropy: Similar to the regression approach, each parameter had particular families of statistics that minimized entropy (Supplementary Documents 8 and 9). Number of alleles, number of shared alleles and FST were commonly in the top ranking, while M-ratio, assignment likelihood and heterozygosity were less common. Importantly, for all parameters except MU, the family of statistics with minimum entropy contained at least one diversity and one differentiation statistic. Typically, the best family also contained statistics that also ranked highly under R^2 (Supplementary Document 6). Usually, minimum entropy was achieved using just two or three of the six families of summary statistics (Supplementary Document 8).

PLSs: The number of PLS components, chosen by graphical examination (Figures 3 and 4), was ~ 10 for each parameter (although sometimes as few as two) for both models as RMSE does not considerably decrease with additional components. Notably, RMSE remained quite high for several parameters (NI and GROWTH, and a lesser extent TB); these parameters also corresponded to those with low R^2 values and are likely difficult to estimate.

Performance in estimating parameters

The mode usually had lower bias and RMSE, and better coverage, than mean and median values (Supplementary Document 10), so the following results focus on performance relating to the mode. Our three methods (MBFV, PLSs and R^2) generally resulted in highly similar performance (that is, if PLSs show high bias, the other methods also do), though very minor differences were apparent (Supplementary Document 11): for example, in the second-wave model, the R^2 method had often the highest factor 2 and lowest

RMSE, while for the null model, the MBFV usually performed best, with the PLS method performing second best in both cases. Here, we present which parameters were estimated well for each model, globally and for the eight local scenarios.

Global (based on 1000 random datasets): Several parameters were estimated well, by multiple measures of performance. At the global level, RMSE for the modal value was ~ 1 or below and bias <0.20 for the null models for GROWTH, MU, NS, TG and TM, and for the second-wave model for GROWTH, MU, NB, NS, TG and TS. The factor 2 was also very good, at or near 0.90 for these parameters. Generally, parameters with poorest estimation were NI and TB. Coverage was very good for all parameters except NI under the null model (Supplementary Document 10).

Local: Overall the worst problems for the second-wave scenario occurred in small and recent introductions (NB10_TS30), in which a high upward bias was seen in some T and N parameters. Overall, the worst problems for the null scenario occurred for small introduction and short bottleneck (NB10_TB2), in which a high upward bias was seen especially in TB and NB. The quality of estimation of each parameter across particular scenarios is explained in much more detail in Supplementary Methods and in Supplementary Document 10.

Application to bumblebee introduction in New Zealand

Posterior probability associated to the single and multiple introduction models for each species are reported in Table 5. Substantial support to the single introduction model was detected for *B. ruderatus* with a probability of 0.74, suggesting that only one of the two documented introductions played a role in the introduction process. For *B. horotum* and *B. terrestris*, the compared models obtained a similar probability and hence it was not possible to discriminate between them.

DISCUSSION

In biological invasions, the genetic variation carried by the invading population arises as a consequence of historical and demographic features of its introduction, including the genetic composition of the source(s), number of introduction events and dynamic of the demographic expansion (Dlugosch and Parker, 2008). In this light, it is fundamental to accurately reconstruct the details of population invasion dynamics, which helps to understand the environmental and evolutionary factors responsible for biological invasions and to facilitate the design of invasion prevention strategies. ABC (Beaumont *et al.*, 2002) methods are increasingly used to draw

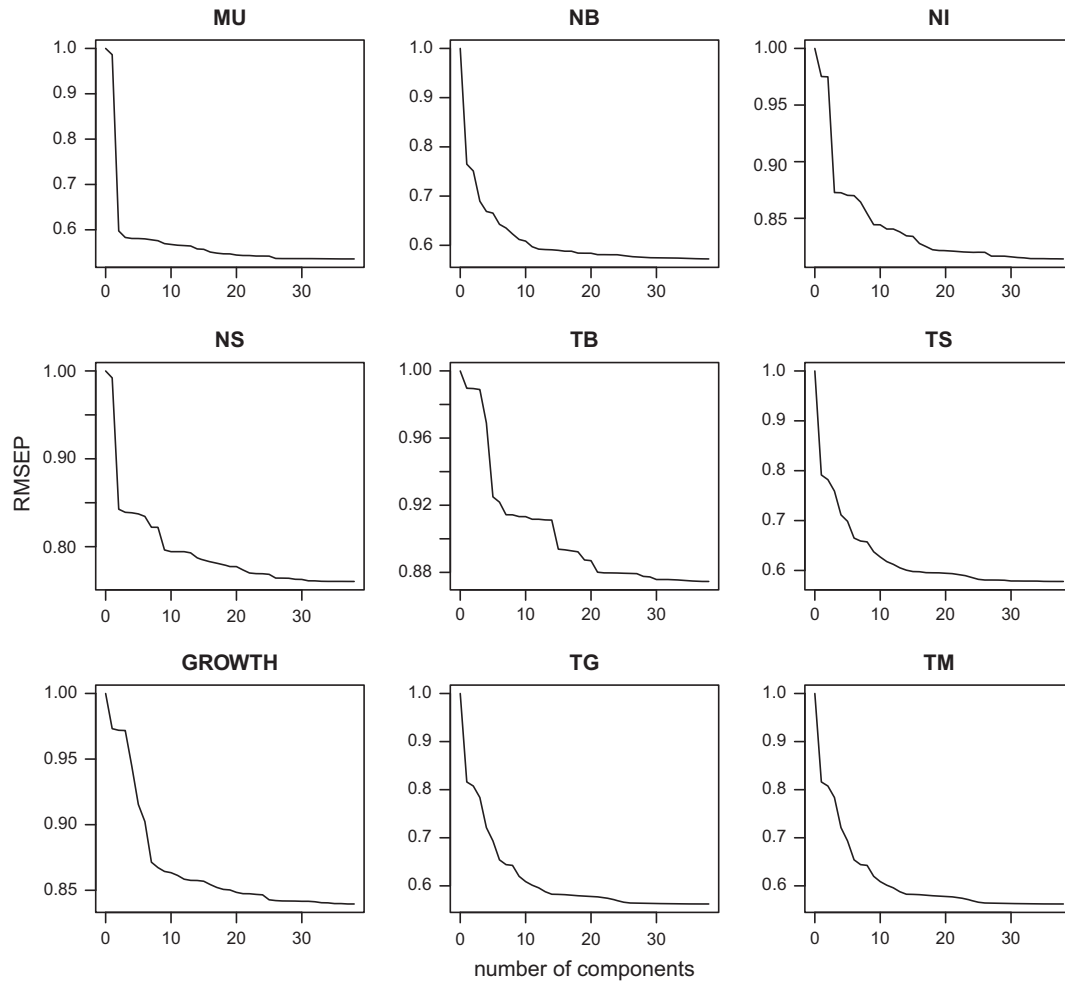


Figure 3 RMSE for up to 38 PLS components for second-wave model.

inferences about the complex evolutionary scenarios typically encountered in the introduction histories of invasive species. This process has been recently favored by the development of user-friendly tools that help to model and statistically compare alternative invasion routes and estimate related parameters (for example, DIYABC, Cornuet *et al.*, 2008), though such software does impose limits on the summary statistics that can be selected and/or on the complexity of the models that can be tested. Therefore, there is an ongoing need to examine the power and reliability of ABC-based methods to correctly choose models and estimate parameters of introduction processes. Examples include Guillemaud *et al.*, 2010, who evaluated the ability of ABC to identify non-sampled (ghost) populations in a multiple-source scenario, and Sousa *et al.*, 2012 who evaluated ABC for quantifying gene flow during population divergence.

In this paper, we proposed a model of invasion accounting for multiple waves of introduction from a single source population, and we evaluated the ability of ABC to correctly recognize the second wave based on simulated microsatellite data. We explicitly compared the null case of a single introduction to a model of two independent waves of migration from the same source. As with any model, the real world is surely more complex than the simulations; however, our goal was to determine, under ideal circumstances of sampling and a simple scenario, if distinguishing a second wave would be possible. If so,

the MISS model may be a valid possibility for future investigators to consider when testing invasion hypotheses on empirical data.

Overall, we found good ability to distinguish a single from a two-wave scenario even with 20 loci, over much but not all of the parameter space tested, which should cover many realistic invasions. Different sets of summary statistics led to only minor differences in power. Further exploration suggested that if the introduction was of small size and long ago (high drift) or large size and with very little time between waves, we were unable, with the ABC procedure, to effectively choose the true model. This situation persisted even using 100 loci, suggesting that when a wrong model is chosen as more likely than the real one, using more loci does not help solving the problem, but rather increases the confidence in the (incorrect) estimates. In the first case, small population size for many generations results in high erosion of diversity and therefore little information content in the genetic data. In the latter case, there is likely insufficient drift after the initial introduction to allow the second wave to contribute a detectable signal.

Nonetheless, the evidence toward the second-wave model when the amount of drift is small was unexpected. We expected that model choice could be difficult, but did not expect strong support to the second-wave model. The erroneous conclusion was likely caused by the effect of specific summary statistics under the ABC procedure. For

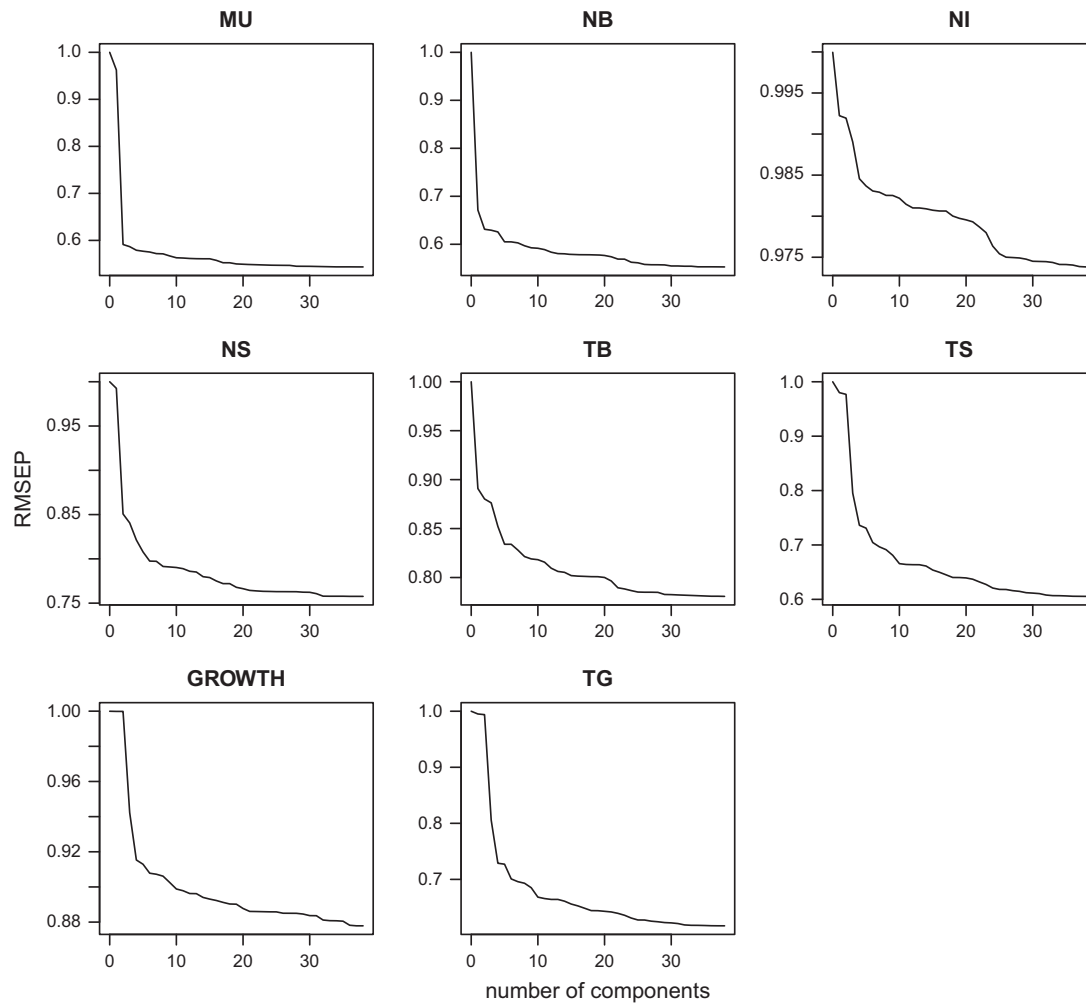


Figure 4 RMSE for up to 38 PLS components for null model.

Table 5 Posterior probabilities supporting the single and multiple introduction models for three *Bombus* species, computed using the polychotomous logistic regression, under the ABC framework using sample sizes indicated in the second and third column

Species	Britain sample size	New Zealand sample size	Probability of single introduction	Probability of multiple introductions
<i>B. ruderatus</i>	28	54	0.74	0.26
<i>B. hortorum</i>	46	37	0.51	0.49
<i>B. terrestris</i>	141	56	0.52	0.48

Abbreviation: ABC, approximate Bayesian computation.

example, the low levels of differentiation (measured by F_{ST}) between the source and the invading population, caused by a short bottleneck period, are easily homogenized by migrants under the second wave model. Specifically, even though the distributions of F_{ST} for the two models tend to be well differentiated (small overlap, see F_{STWC} , Supplementary Figure 1), they were both at extremely low F_{ST} values. The same behavior was detected for other differentiation statistics, number of shared/exclusive alleles and expected heterozygosity. Thus, summary statistics seem to confound the two models under the ABC model choice procedure when the amount of drift is low. This result confirmed the utility of the ABC inferential framework to distinguish

one from two waves of colonization on the condition that the amount of time between the two waves is not too small (at least 5–10 generations). Usefully, simulations at $TB=5$ and $TB=10$ showed that second waves with small introductions and bottlenecks of moderate length can be correctly detected. A stricter decision threshold also helped in some problematic situations but is not a universal solution.

Most parameters were well estimated globally, whether simulated data was created under the null (single wave) or two-wave model. The most problematic parameter was TB (length of bottleneck, high upward bias), followed by NI (current invasion size). In exploring the local scenarios, it was apparent that poor estimation of TB is

primarily in cases of both small introductions and short bottlenecks. It is unsurprising that NI is also a difficult parameter—the current large size of the invading population and its recent expansion may leave little genetic signal. The parameters of highest practical interest, the timing and size of the introductions (TS and NB), were most reliably estimated for moderate to large introductions and either moderate to long bottlenecks or older introductions. In these situations, estimates always had RMSE < 1 and are nearly always within a factor of 2 of the true value. This performance should be sufficient for some management applications.

We applied ABC to explicitly compare the MISS model to a single introduction event for the three *Bombus* species introduced to New Zealand a century ago, in two document events. For two species, we were unable to support either scenario. This may be due to the lack of genetic information contained in eight loci; power to identify the right model might be improved by increasing the number of loci to ≥ 20 . For a third species, *B. ruderatus*, we clearly supported a single introduction event. Since more than 20 generations separate the two waves, the amount of drift should be sufficient to discriminate among models and hence avoid strong support toward the wrong model. From this result we can infer that even though *B. ruderatus* may have been present among the assorted bumblebees introduced in 1885 and 1906, in fact only one of these introductions successfully contributed to the invasion for this species. As Lye *et al.*, 2011 reported, mortality was likely high for the bumblebee introduction, and thus it may not be surprising that only one of the introductions contributed to the invasion.

Comparison to past work

ABC analyses of invasion scenarios are increasingly sophisticated and complex, including the use of a nested approach to compare potentially dozens of models (Boissin *et al.*, 2012; Boubou *et al.*, 2013; Konečný *et al.*, 2013), some of which feature admixture, ghost populations and other features. However, the MISS scenario is rarely, if ever, considered as a competing model in invasions. In some situations MISS will not be a relevant model to consider, but the mention of a possible second wave in some investigations (see Introduction section) suggests it sometimes warrants serious consideration. In the few discussions of the second-wave model that we are aware of (see Introduction section), investigators have postulated that distinguishing a second wave from a single-wave model may not be possible. We demonstrated otherwise; across a range of parameter values, the second wave was distinguishable, and several parameters were well estimated using 100 or 20 microsatellites. Notably, the number of individuals in each wave was well estimated. This is a particularly important parameter for management decisions such as eradication or detection (Ficetola *et al.*, 2008).

It is notable that we did not observe a ‘best’ set of summary statistics for model choice using either intuition or previous methods (Veeramah *et al.*, 2012). Different subsets of statistics only minimally altered power in most cases. Importantly, we observed good results using all statistics (no issues of dimensionality) for model choice or estimating parameters. In MISS, and possibly in other ABC investigations, choosing statistics whose distributions show least overlap (KS test) in the opposing models may not help identify the ‘best’ summary statistics; these statistics may not necessarily contain the most information about the processes occurring in the model. In an investigation of early divergence of modern humans in Africa, Veeramah *et al.*, 2012, using the KS *P*-value ranking procedure we used, found that various sets of summary statistics all provided quite similar power, and in fact a single summary statistic provided the

highest power. We observed that in two of our scenarios, the smallest set (five statistic) provided increased power over the other three sets; this was balanced, however, by an increase in error rates for the null model PODs. It is possible that, as demonstrated by the few families in the minimum entropy sets, two to four summary statistics may contain nearly all of the relevant information, but small sets could also produce misleading signals.

Summary statistics for model choice are clearly still a major challenge for ABC work. Still, for our situation, minimum entropy sets (normally containing some combination of number alleles, heterozygosity, *F*_{ST} and number shared alleles) performed nearly as well as PLS or *R*² methods; differences between the three methods were all quite small. We recommend always to use at least one diversity-based combined with one differentiation-based statistic. Furthermore, statistics calculated on the invasive population seem to contain most of the useful information. Lastly, we note that sometimes top *R*² statistics included those that are little used in ABC, such as Jost’s *D*. Future MISS investigators should probably test an array of summary statistics and various combinations.

Additional remarks and caveats

It may be important to distinguish one from two waves of colonization from a single source for several reasons. A second wave may provide demographic or genetic reinforcement to the initial colonists, or help relieve inbreeding among founders (Lodge *et al.*, 2006). Such reinforcement has been suggested as a possible reason for the lag time often observed in invasions (Lee 2002; Boubou *et al.*, 2013). Knowledge of whether second (or third and so on) waves occurred in invading populations can contribute to understanding evolutionary and ecological dynamics of invasions. In the MISS model, it is explicitly recognized that a second wave may come from the same, rather than a geographically distinct, source to provide reinforcement. From a practical perspective, detecting a second wave and quantifying the number of individuals in each wave has implications for monitoring and management, which are most effective if tailored to the invading species, propagule size, source and so on (Estoup and Guillemaud 2010; Guillemaud *et al.*, 2011). If second waves are common, managers could use this information to decide whether, how and when to isolate or eliminate small introduced populations, for example, before subsequent waves occurs (Dlugosch and Parker, 2008). There are also implications for inspection and interception; if second waves commonly occur, strict inspection would be important to continue even after first establishment to prevent further waves.

We only considered a simple situation of no admixture with other invading populations, no spatial expansion and so on. Indeed, two important assumptions of the MISS model are that meaningful levels of migration do not occur between the focal invasive population and other invasive populations, and the two waves are of similar size; whether these assumptions are met in real invasions is not currently well known. Our model may be most directly applicable to cases of crop pest outbreaks, introduction of biological controls or rare inter-continental introductions in which MISS might be a likely mechanism (Miller *et al.*, 2005; Lye *et al.*, 2011). In addition, the MISS model might be a component of more complex models of invasion, though migration from another population (especially if it occurs early in the invading population’s history) could possibly swamp signatures of the second wave, if one occurred. This hypothesis, and the ability to detect a second wave in complex situations of later admixture or multiple sources, will need to be tested.

A final point is that we attempted to estimate all parameters including those regarding timing (such as TS, time of the split), with

quite broad priors. In some real invasions, timing or other parameters may be known with a small degree of error. Inclusion of such information as priors in empirical ABC studies (Estoup *et al.*, 2004 and 2010) may likely improve both model selection and parameter estimation for MISS, over the already good results we observe with broad priors.

CONCLUSION

With this work, we demonstrated that it is often possible, using the inferential power of ABC and under a range of realistic demographic conditions, to distinguish between one or two waves of invasion from the same source, and to infer some important parameters. We suggest that it may be fruitful to consider the MISS model in future investigations of biological invasions.

DATA ARCHIVING

Original data associated with this article (including ABC reference tables, input files for recreating simulations, a power analysis script and summary statistics from the *Bombus* samples) are archived in Dryad: doi:10.5061/dryad.2c20k. Our study also used datasets from Lye *et al.*, 2011 found in Dryad: doi:10.5061/dryad.sk22v.

AUTHOR CONTRIBUTIONS

AB originated the MISS concept and performed simulations and ABC. STV searched the literature for realistic parameters. SH, SG and AB analyzed data. SH drafted the manuscript. All authors discussed results and revised the manuscript, and together planned the study and analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Giorgio Bertorelle and Guido Barbujani for valuable suggestions, and Francesca Tassi for helping with figures. STV was supported by a PhD Fellowship for International students at the University of Ferrara. AB, SH and SG were supported by the University of Ferrara. SH was additionally supported by European Commission grant FP7-ENV-2009-1-244250, which established the ConGRESS (Conservation Genetic Resources for Effective Species Survival) project. This work was completed while SH was a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from the University of Tennessee, Knoxville.

DATA ACCESSIBILITY

Our simulation parameter files, reference tables and representative datasets will be archived at Dryad.

- Bertorelle G, Benazzo A, Mona S (2010). ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* **19**: 2609–2625.
- Besnard G, Dupuy J, Lareter M, Cuneo P, Cooke D, Chikhi L (2014). History of the invasive African olive tree in Australia and Hawaii: evidence for sequential bottlenecks and hybridization with the Mediterranean olive. *Evol Appl* **7**: 195–211.
- Boissin E, Hurlley B, Wingfield MJ, Vasaitis R, Stenlid J, Davis C *et al.* (2012). Retracing the routes of introduction of invasive species: the case of the *Sirex noctilio* woodwasp. *Mol Ecol* **21**: 1–17.
- Boubou A, Migeon A, Roderick GK, Auger P, Cornuet JM, Magalhães S *et al.* (2013). Test of colonisation scenarios reveals complex invasion history of the red tomato spider mite *Tetranychus evansi*. *PLoS One* **7**: e35601.
- Clegg SM, Degnan SM, Kikkawa J, Moritz C, Estoup A, Owens IP (2002). Genetic consequences of sequential founder events by an island-colonizing bird. *Proc Natl Acad Sci* **99**: 8127–8132.
- Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ *et al.* (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713–2719.
- Csilléry K, Blum MGB, Gaggiotti OE, François O (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**: 410–418.
- Diez JM, D'Antonio CM, Dukes JS *et al.* (2012). Will extreme climatic events facilitate biological invasions?. *Front Ecol Environ* **10**: 249–257.
- Dlugosch KM, Parker IM (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol Ecol* **17**: 431–449.
- Estoup A, Baird SJE, Ray N, Currat M, Cornuet JM, Santos F *et al.* (2010). Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad *Bufo marinus*. *Mol Ecol Resour* **10**: 886–901.
- Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution* **58**: 2021–2036.
- Estoup A, Guillemaud T (2010). Reconstructing routes of invasion using genetic data: why, how and so what? *Mol Ecol* **19**: 4113–4130.
- Excoffier L, Estoup A, Cornuet J-M (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.
- Excoffier L, Foll M (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**: 1332–1334.
- Excoffier L, Lischer HE (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–567.
- Ficetola GF, Bonin A, Miao C (2008). Population genetics reveals origin and number of founders in a biological invasion. *Mol Ecol* **17**: 773–782.
- Fonseca DM, Widdel AK, Hutchinson M, Spichiger SE, Kramer LD (2010). Fine-scale spatial and temporal population genetics of *Aedes japonicus*, a new US mosquito, reveal multiple introductions. *Mol Ecol* **19**: 1559–1572.
- Garza JC, Williamson EG (2001). Detection of reduction in population size using data from microsatellite loci. *Mol Ecol* **10**: 305–318.
- Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* **104**: 88–99.
- Guillemaud T, Ciosi M, Lombaert E, Estoup A (2011). Biological invasions in agricultural settings: Insights from evolutionary biology and population genetics. *C R Biol* **334**: 237–246.
- Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**: 409–417.
- Keller SR (2009). Adaptation and colonization history affect the evolution of clines in two introduced species. *New Phytol* **183**: 678–690.
- Konečný A, Estoup A, Duplantier J-M, Bryja J, Bâ K, Galan M *et al.* (2013). Invasion genetics of the introduced black rat (*Rattus rattus*) in Senegal, West Africa. *Mol Ecol* **22**: 286–300.
- Lankau RA, Nuzzo V, Spyreas G, Davis AS (2009). Evolutionary limits ameliorate the negative impact of an invasive plant. *Proc Natl Acad Sci* **106**: 15362–15367.
- Lee CE (2002). Evolutionary genetics of invasive species. *Trends Ecol Evol* **17**: 386–391.
- Lodge DM, Williams SL, MacIsaac H, Hayes KR, Leung B, Reichard S *et al.* (2006). Biological invasions: recommendations for US policy and management. *Ecol Appl* **16**: 2035–2054.
- Lombaert E, Guillemaud T, Cornuet J, Malausa T, Facon B, Estoup A *et al.* (2012). Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *PLoS One* **5**: e9743.
- Lye GC, Lepais O, Goulson D (2011). Reconstructing demographic events from population genetic data: the introduction of bumblebees to New Zealand. *Mol Ecol* **20**: 2888–2900.
- Marsico TD, Wallace LE, Ervin GN, Brooks CP, McClure JE, Welch ME (2010). Geographic patterns of genetic diversity from the native range of *Cactoblastis cactorum* (Berg) support the documented history of invasion and multiple introductions for invasive populations. *Biol Invasions* **13**: 857–868.
- Mevik B-H, Wehrens R (2007). The PLS package: principal component and partial least squares regression in R. *J Stat Software* **18**: 1–24.
- Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, Derridj S *et al.* (2005). Multiple transatlantic introductions of the western corn rootworm. *Science* **310**: 992.
- Muirhead JR, Gray DK, Kelly DW, Ellis SM, Heath DD, MacIsaac HJ (2008). Identifying the source of species invasions: sampling intensity vs genetic diversity. *Mol Ecol* **17**: 1020–1035.

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

Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005). Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21**: 1733–1734.

Beaumont MA (2008). Joint determination of topology, divergence time and immigration in population trees. In: Matsumura S, Forster P, Renfrew C (eds) *Simulations, Genetics and Human Prehistory*. Cambridge: McDonald Institute Monographs.

Beaumont MA (2010). Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst* **41**: 379–406.

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

- Neuenschwander S, Lurgiàdèr CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008). Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**: 757–772.
- Nunes MA, Balding DJ (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Stat Appl Genet Mol Biol* **9**: 1–14.
- Olsen KM, Gross BL (2008). Detecting multiple origins of domesticated crops. *Proc Natl Acad Sci* **105**: 13701–13702.
- Paetkau D, Calvert W, Stirling I, Strobeck C (1995). Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* **4**: 347–354.
- Pascual M, Chapuis MP, Mestres F, Balanyà J, Huey RB, Gilchrist GW *et al.* (2007). Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Mol Ecol* **16**: 3069–3083.
- Phillips BL, Brown GP, Webb JK, Shine R (2006). Invasion and the evolution of speed in toads. *Nature* **439**: 803.
- Sousa VC, Beaumont M a, Fernandes P, Coelho MM, Chikhi L (2012). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity* **108**: 521–530.
- Stocks M, Siol M, Lascoux M, De Mita S (2014). Amount of information needed for model choice in approximate Bayesian computation. *PLoS One* **9**: e99581.
- Tenenhaus M, Gauchi J-P, Ménardo C (1995). Regression PLS and applications. *Rev Stat Appl* **43**: 7–63.
- Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G *et al.* (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol Biol Evol* **29**: 617–630.
- Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinform* **11**: 116.
- Wilson JRU, Dormontt EE, Prentis PJ, Lowe AJ, Richardson DM (2008). Something in the way you move: dispersal pathways affect invasion success. *Trends Ecol Evol* **24**: 136–144.

Supplementary Information accompanies this paper on Heredity website (<http://www.nature.com/hdy>)