MOLECULAR ECOLOGY

Molecular Ecology (2011) doi: 10.1111/j.1365-294X.2011.05322.x

Inferring the origin of populations introduced from a genetically structured native range by approximate Bayesian computation: case study of the invasive ladybird *Harmonia axyridis*

E. LOMBAERT,* T. GUILLEMAUD,* C. E. THOMAS,† L. J. LAWSON HANDLEY,† J. LI,‡ S. WANG,§ H. PANG,¶ I. GORYACHEVA,** I. A. ZAKHAROV,** E. JOUSSELIN,†† R. L. POLAND,‡‡ A. MIGEON,†† J. VAN LENTEREN,§§ P. DE CLERCQ,¶¶ N. BERKVENS,¶¶ W. JONES*** and A. ESTOUP††

**INRA, UMR 1301 IBSV (INRA/Université de Nice Sophia Antipolis/CNRS), 400 Route des Chappes, BP 167-06903 Sophia Antipolis Cedex, France, †University Hull, Evolutionary Biology Group, Kingston-Upon-Hull HU6 7RX, N Humberside, UK, ‡College of Environment and Plant Protection, Hainan University, Haikou 520228, China, §Group of Invasive Species Identification and Management, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, ¶School of Life Sciences, State Key Laboratory of Biocontrol, Sun Yat-Sen University, Guangzhou 510275, China, **Vavilov Institute of General Genetics, Moscow State University, Moscow, Russia, ††INRA, UMR CBGP (INRA, IRD, Cirad, Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez Cedex, France, ‡‡Clifton College, 32 College Road, Clifton, Bristol BS8 3JH, UK, §§Laboratory of Entomology, Wageningen University, 6708PB Wageningen, The Netherlands, ¶¶Department of Crop Protection, Ghent University, Coupure Links 653, B-9000 Ghent, Belgium, ***Biological Control of Pests Research, National Biological Control Laboratory, ARS, US, DA, PO Box 67, Stoneville, MS, USA*

Abstract

Correct identification of the source population of an invasive species is a prerequisite for testing hypotheses concerning the factors responsible for biological invasions. The native area of invasive species may be large, poorly known and/or genetically structured. Because the actual source population may not have been sampled, studies based on molecular markers may generate incorrect conclusions about the origin of introduced populations. In this study, we characterized the genetic structure of the invasive ladybird *Harmonia axyridis* in its native area using various population genetic statistics and methods. We found that native area of *H. axyridis* most probably consisted of two geographically distinct genetic clusters located in eastern and western Asia. We then performed approximate Bayesian computation (ABC) analyses on controlled simulated microsatellite data sets to evaluate (i) the risk of selecting incorrect introduction scenarios, including admixture between sources, when the populations of the native area are genetically structured and sampling is incomplete and (ii) the ability of ABC analysis to minimize such risks by explicitly including unsampled populations in the scenarios compared. Finally, we performed additional ABC analyses on real microsatellite data sets to retrace the origin of biocontrol and invasive populations of *H. axyridis*, taking into account the possibility that the structured native area may have been incompletely sampled. We found that the invasive population in eastern North America, which has served as the bridgehead for worldwide invasion by *H. axyridis*, was probably formed by an admixture between the eastern and western native clusters. This admixture may have facilitated adaptation of the bridgehead population.

Keywords: biocontrol, biological invasion, harlequin ladybird, invasive species, microsatellite, source population

Received 18 March 2011; revision received 15 September 2011; accepted 16 September 2011

Correspondence: Eric Lombaert, Fax: +33 4 92 38 64 01;
E-mail: lombaert@sophia.inra.fr

Introduction

Historical and observational data for invasive species are often sparse and incomplete, so molecular genetic markers are increasingly used and have proved to be efficient tools for the inference of invasion routes (Estoup & Guillemaud 2010). However, such inference remains a major challenge, because of two specific features of invasions. First, invasion history is often marked by stochastic genetic and demographic events, which may make it difficult to interpret the observed genetic patterns. In particular, introduction is often characterized by a loss of genetic diversity relative to the source population (a founder event) and may be followed by a demographic bottleneck, resulting in strong genetic drift and substantial genetic differentiation between the introduced population and all other populations, including the source population. Moreover, multiple introductions may give rise to genetic admixtures between several differentiated populations in the invasive range, thus generating unique genetic combinations that are not found together in the native range. Second, sampling issues may compromise inference. An invasive population may be derived from different types of source population: (i) populations from the native area that may be large, poorly known and/or genetically structured or (ii) other invasive outbreak(s), which serve as a source of colonists for other areas, the existence of which may be unknown because they occur in unexpected or unexplored areas. The actual geographical range of a target species may be large and difficult to explore exhaustively and, in many cases, the actual source population may not have been sampled.

The use of approximate Bayesian computation (ABC, Beaumont *et al.* 2002; Bertorelle *et al.* 2010; Csillery *et al.* 2010) on molecular data makes it possible to generate model-based inferences for complex scenarios, such as those related to the introduction histories of invasive species (Estoup & Guillemaud 2010). This method has recently been successfully used to retrace the invasion routes of various invasive species, explicitly taking into account demographic and genetic stochasticity resulting from bottlenecks, multiple introductions and/or genetic admixture events (Miller *et al.* 2005; Pascual *et al.* 2007; Lombaert *et al.* 2010). Studies of simulated data have shown that, in most cases, ABC is more powerful in this context than other more traditional methods for population genetics studies, such as neighbour-joining trees or *F*-statistics (Estoup & Guillemaud 2010; Guillemaud *et al.* 2010; Lombaert *et al.* 2010).

Another advantage of the ABC method is that it allows the explicit inclusion of unsampled populations in the evolutionary scenarios compared, although the power of ABC to deal with unsampled populations has been little investigated (but see Guillemaud *et al.* 2010). The native range of a species is characterized by a long evolutionary history shaped by mutation, drift, migration and selection operating in a spatially and temporally heterogeneous environment. A strong geographical genetic structure is therefore often found in the native range of invasive species (e.g. Kolbe *et al.* 2004; Ciosi *et al.* 2008). Exhaustive sampling is difficult in native areas that are often large and may be poorly known. It is therefore important to evaluate the effects of unsampled native source populations on the inference of introduction routes in the presence of genetic structure within the native area. We addressed this question with controlled simulated microsatellite data sets and real data sets obtained from wild and biocontrol populations of the harlequin ladybird *Harmonia axyridis*.

The native area of *H. axyridis* covers a large part of Asia (Kazakhstan, southern Siberia, Mongolia, eastern China, Korea and Japan, reviewed in Poutsma *et al.* 2008). *H. axyridis* has been repeatedly introduced into North America since 1916 as a biocontrol agent for aphids. Several source populations are known to have contributed to American biocontrol stocks, including, in particular, the populations of Eastern Siberia, China, South Korea and Japan (Tedders & Schaefer 1994; Krafsur *et al.* 1997). In Europe, biocontrol with *H. axyridis* began in the early 1990s, with individuals derived from a single population brought from China in 1982 by an INRA laboratory (Ongagna *et al.* 1993), which was subsequently reared in research laboratories and several biofactories. This same European biocontrol population was also used repeatedly in South America from 1986 (Argentina and Brazil Poutsma *et al.* 2008). Despite the recurrent intentional releases of ladybirds for acclimation attempts in Europe and South America, the species took decades to establish itself (Koch 2003). However, for unknown reasons, it recently suddenly became highly invasive on four continents. Invasive populations were first recorded in eastern (Louisiana, USA, Chapin & Brou 1991) and western (Oregon, USA, LaMana & Miller 1996) North America in 1988 and 1991, respectively. They were then recorded in Europe (Belgium, Adriaens *et al.* 2003) and South America (Argentina, Saini 2004) in 2001 and in Africa (South Africa, Stals & Prinsloo 2007) in 2004. The species has widely spread in these areas and has become a major predator of nontarget arthropods, a household invader, and a pest in fruit crops (Koch 2003). Using ABC methods on microsatellite and historical data, Lombaert *et al.* (2010) showed that the two North American outbreaks originated from two independent introductions from the native area, but the exact geographical origins of the source populations were not investigated. They also found that the eastern North American (ENA) population acted as a bridgehead for worldwide invasion, acting as the source population of the European, South American and African outbreaks, with some admixture with the biocontrol population in Europe.

In this study, we characterized the genetic structure of *H. axyridis* in its native area by Bayesian clustering methods and more classical population genetic statistics and methods (e.g. F_{ST} and neighbour-joining trees). We then performed ABC analyses on controlled simulated microsatellite data sets to evaluate (i) the risk of selecting incorrect scenarios when using an incomplete sampling strategy in a genetically structured native area and (ii) the ability of ABC analysis to minimize such risks by explicitly including unsampled populations in the scenarios compared. Finally, we performed additional ABC analyses to retrace the origin of biocontrol and invasive populations of *H. axyridis*, taking into account the possibility that the structured native area may have been incompletely sampled.

Methods

Sampling and genotyping

Harmonia axyridis samples were collected within the native area, at nine sites, covering a substantial part of the natural distribution of this species (Kazakhstan, Russia, China, South Korea and Japan; Fig. 1; Table S1, Supporting information). Three of these samples were previously used by Lombaert *et al.* (2010). We also collected five European biocontrol samples believed to be derived from the original 1982 INRA sample. Three of these samples were obtained from different commercial biofactories, and two were obtained from INRA rearing stocks from 1987 and 2006 (Table S1, Supporting information). The oldest INRA sample, EB-INRA87, corresponds to that used by Lombaert *et al.* (2010). A large number of native populations have been used for biocontrol in North America (Tedders & Schaefer 1994; Krafsur *et al.* 1997; Koch 2003), but only one sample, collected in 1980, could be obtained and analysed (http://www.ars-grin.gov/cgi-bin/nigrp/robo/f941spl?50902) (Table S1, Supporting information). The samples representative of the five invaded areas described by Lombaert *et al.* (2010) were also used in the present study (Table S1, Supporting information). The sample size for each population ranged from 18 to 42 individuals (mean 29.3; Table S1, Supporting information). Samples were genotyped at 18 microsatellite markers, as described by Loiseau *et al.* (2009). Four biocontrol populations were obtained from insect collections and had been stored dry at room temperature for a long period of time, greater than 20 years in some cases (Table S1, Supporting information). The DNA extracted from these samples was highly degraded, necessitating several modifications to the protocols described by Loiseau *et al.* (2009): (i) DNA was extracted from entire bodies (rather than just the pronotum and head), (ii) annealing temperature for PCR was set at 55 °C (rather than 57 °C) and (iii) the number of PCR cycles was set at 35 (rather than 25).

Genetic variation within and between populations

Genetic variation within samples was quantified by calculating the mean expected heterozygosity H_e (Nei 1987) and the mean allelic richness (AR) with the rarefaction method of Leberg (2002) in FSTAT (version 2.9.3.2 Goudet 2002). Amplification was difficult at eight of the 18 microsatellite loci in the four biocontrol samples that had been stored dry (i.e. allelic PCR profiles could be safely interpreted for only a small number of individuals for these eight loci). We therefore calculated two indices of allelic richness: AR_{10} was calculated for 10 microsatellite loci for all 20 populations, whereas AR_{18} was calculated for the entire set of 18 microsatellite data for a subset of 16 populations.

Genetic variation between populations was summarized by calculating pairwise F_{ST} estimates as described by Weir & Cockerham (1984), with Genepop (Raymond & Rousset 1995b). Exact tests for population genotypic differentiation (Raymond & Rousset 1995a) were carried out for all pairs of populations within the native area, with the same software. Because these tests involve nonorthogonal and multiple comparisons, we corrected significance levels by the false discovery rate procedure (Benjamini & Hochberg 1995). We plotted a neighbour-joining (NJ) tree (Saitou & Nei 1987), using the pairwise genetic distances described by Cavalli-Sforza and Edwards (Cavalli-Sforza & Edwards 1967), in Populations 1.2.30 software (http://bioinformatics.org/~tryphon/populations/). The robustness of tree topology was evaluated by carrying out 1000 bootstrap replicates over loci.

Population structure and isolation by distance within the native area

The clustering approach implemented in STRUCTURE v2.3.3 (Pritchard *et al.* 2000) was used to infer the number of potential population units within the native area of *H. axyridis*. We chose the admixture model with correlated allele frequencies and, because our sampling scheme involved the collection of many individuals from a few discrete distant locations (Schwartz & McKelvey 2009), we used the sampling location as prior

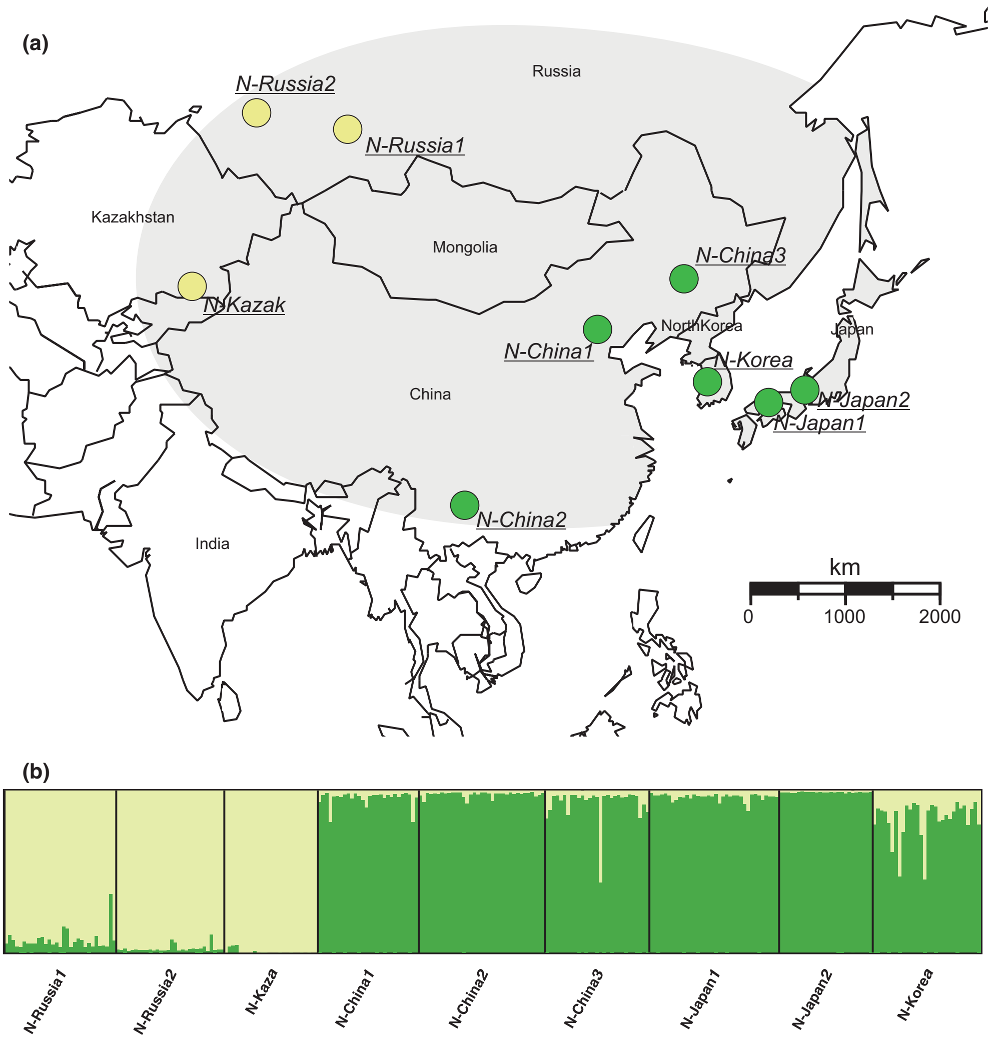

Fig. 1 Geographical origins and genetic clustering of sampled native populations of the Asian ladybird *Harmonia axyridis*. (a) Locality codes are underlined and in italics (see Table S1, Supporting information for details about the sites sampled). The shaded area approximately corresponds to the known native distribution of the species. Sampled sites with similar colours belong to the same genetic cluster, as assessed by the spatial group clustering method of Corander *et al.* (2004) implemented in BAPS software. (b) Ancestry estimation based on the Bayesian clustering method STRUCTURE in the native *Harmonia axyridis* samples, assuming two population clusters ($K = 2$). Each vertical line represents an individual, and each colour represents a cluster. Individuals are grouped by sampling location (at the bottom).

information (Hubisz *et al.* 2009). We used default values for all other parameters of the software. Each run consisted of a burn-in period of 10^5 Markov chain Monte Carlo (MCMC) iterations, followed by 10^6 MCMC iterations. We carried out 20 replicate runs for each value of the number (K) of clusters, set between 1 and 9 (i.e. the number of samples). The natural logarithm of the likelihood of the data $\ln(P(X|K))$ was calculated: it is expected to be high with a low variance for the true K (Pritchard *et al.* 2000).

We also used the clustering approach based on groups of individuals (i.e. population samples) implemented in BAPS 5.2 software (Corander *et al.* 2004), with the spatial coordinates of the samples as prior information. We conducted a series of 20 replicate runs, with the upper limit for the number of clusters set at 9 (the actual number of sampled native sites) for each run.

Finally we tested for isolation by distance patterns within the native range. The model of isolation by distance predicts that the genetic distances between populations, as measured by pairwise $F_{ST}/(1-F_{ST})$, increase approximately linearly with logarithm of spatial distances (Rousset 1997). We conducted this method for (i) the whole set of samples throughout the native area and (ii) for the genetic clusters inferred by the aforementioned clustering approaches when the number of sites sampled within a cluster was sufficient. All the correlations between the natural logarithmic distances and the pairwise $F_{ST}/(1-F_{ST})$ were tested using Mantel tests with 10 000 permutations on the Spearman's rank correlation coefficient as implemented in SPAGeDI (Hardy & Vekemans 2002).

ABC analyses on controlled simulated data sets

We ran computer simulations to investigate the impact of genetic structure in the native range on the ability to determine the origin of an introduced population by ABC. All simulations of data sets and ABC analyses were performed with DIYABC v.1 (Cornuet *et al.* 2010). We focused on the simple case of a native range structured into two main population clusters, cluster A and cluster B (the situation identified for *H. axyridis*, see the Results section), within which substructure could exist. We considered that one population was sampled from each of the two native clusters (native sample A and native sample B) and one invasive population, whose origin depends on the scenario considered. Following our sampling design for real *H. axyridis* samples, the number of diploid individuals was fixed at 30 in each of the three population samples, and the data set consisted of genotypes for 18 statistically independent microsatellite loci. No migration between any pair of populations was assumed. Two sets of scenarios were devised, with three competing scenarios in each set:

The sampled origin scenario set (SO): the invasive population originated directly from one of the two sampled native populations (scenarios SA and SB for clusters A and B, respectively; Fig. 2) or from an admixture of these two populations (scenario SAB; Fig. 2).

The unsampled origin scenario set (UO): we simulated substructuring within each native cluster. The introduced population originated from an unsampled population belonging to cluster A or B (scenarios UA and UB for clusters A and B, respectively; Fig. 2) or from an admixture of the two unsampled populations (scenario UAB; Fig. 2).

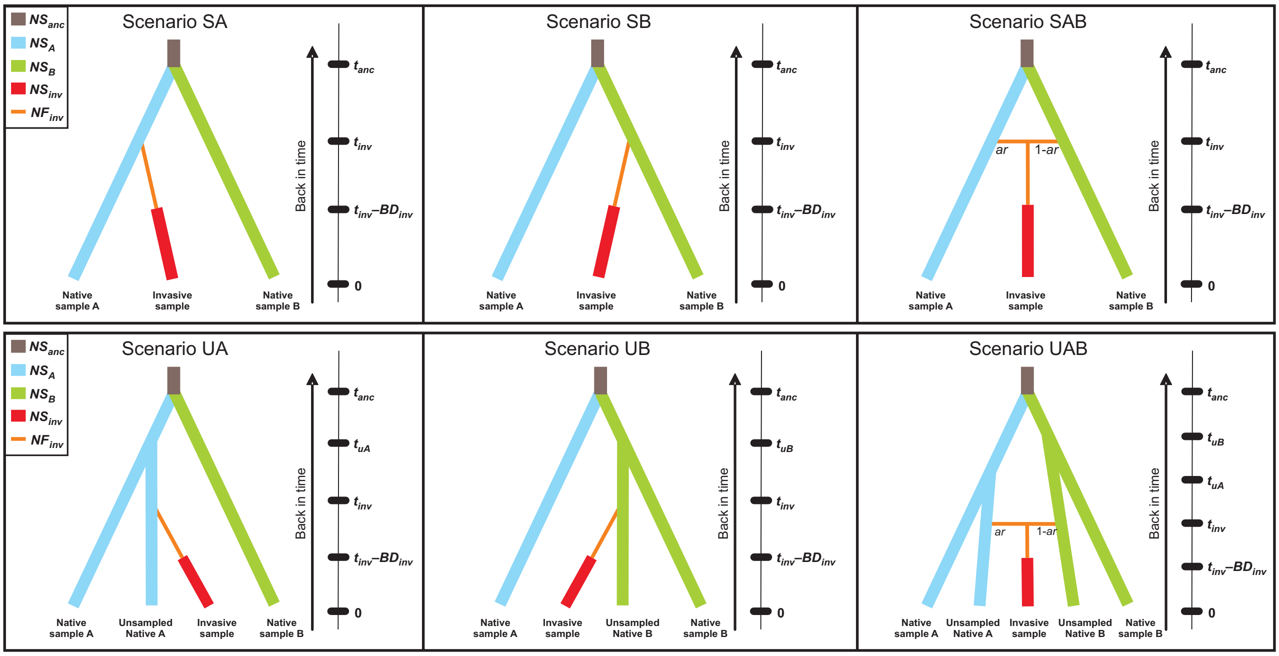

Fig. 2 Graphical representation of the two sets of competing scenarios used for the ABC analyses on controlled simulated data sets. Scenarios SA, SB and SAB correspond to the sampled origin scenario set (SO), and scenarios UA, UB and UAB correspond to the unsampled origin scenario set (UO). Historical and demographic parameters were the same for all introduction models. Time 0 is the sampling date. The invasive population was founded t_{inv} generations ago, had an effective number of founders, NF$_{\mathrm{inv}}$, remaining constant for a few generations (bottleneck duration BD$_{\mathrm{inv}}$) and then reached a larger stable effective population size, NS$_{\mathrm{inv}}$. The two native clusters, A and B, merged in an ancestral unsampled population t_{anc} generations ago. Effective population sizes were stable over time and equal to NS$_{\mathrm{A}}$, NS$_{\mathrm{B}}$ and NS$_{\mathrm{anc}}$ in the populations of clusters A, B and the ancestral population, respectively. When admixture occurs, the admixture rate *ar* is the genetic contribution of the native population from cluster A. In the unsampled origin scenario set (scenarios UA, UB and UAB), each unsampled native population merges into the sampled native population at time t_{uA} and t_{uB} (for clusters A and B, respectively), with t_{uA} and $t_{\mathrm{uB}} \geq t_{\mathrm{inv}}$ and $\leq t_{\mathrm{anc}}$. For all models, populations were assumed to be isolated from each other, with no exchange of migrants. All parameters with associated distributions are described in Table 1.

ABC analysis was performed with historical, demographic and mutational parameter values drawn from the prior distributions described in Table 1 ('broad parameter distribution set') and by simulating two reference tables (i.e. set of summary statistics computed from data simulated according to each model, with parameters drawn from the prior distributions), one based on the three sampled origin (SO) scenarios and the other on the three unsampled origin (UO) scenarios. Each reference table contains 10^6 simulated microsatellite data sets per scenario. We summarized the genetic variation within and between populations, using a set of statistics that we successfully employed in previous ABC analyses (Cornuet *et al.* 2008; Guillemaud *et al.* 2010; Lombaert *et al.* 2010). For each population and each population pair, we used the mean number of alleles per locus, the mean expected heterozygosity (Nei 1987) and the mean allelic size variance. The other statistics used were the mean ratio of the number of alleles over the range of allele sizes (Garza & Williamson 2001), the pairwise F_{ST} values (Weir & Cockerham 1984), the mean individual assignment likelihoods of population *i* being assigned to population *j* and the maximum likelihood estimate of admixture proportion (Pascual *et al.* 2007). Overall, a total number of 31 summary statistics was used. All these statistics are thought to be informative in this study. Both lack and excess of summary statistics can be troublesome for model selection. Unfortunately, there is still no general rule or method as to which and how many summary statistics should be used in an ABC analysis. Recent improvements of ABC get round this problem using dimension reduction techniques, including a nonlinear feed-forward neural network (Blum & Francois 2010) and partial least squares regression (Wegmann *et al.* 2009). These types of algorithms have not been implemented yet in the DIYABC package. The added value of such algorithms in the context of complex models and large data sets remains, however, to be thoroughly tested (Bertorelle *et al.* 2010; But see Hamilton *et al.* 2005; Joyce & Marjoram 2008; Nunes & Balding 2010). Most importantly, it is worth stressing that the aforementioned dimension reduction techniques have been developed mostly for the estimation of posterior distribution of demographic parameters under a given scenario and not for the discrimination among a set of competing scenarios. Our set of statistics may not be optimal, which may reduce the ability of finding the true scenario, but we believe that we will still be able to properly compare the power of the UO and SO scenario sets.

For each of the six scenarios described previously, we simulated pseudo-observed genetic data sets (referred to hereafter as 'pods') with parameters drawn either from the same distributions as the large prior distributions (Table 1, 'broad parameter distribution set') or from an alternative narrower set of distributions mimicking the low level of differentiation and high level of diversity found within the native area of *H. axyridis* (see Results, Table 1, 'HA-like parameter distribution set', Table S2 and Fig. S1, Supporting information). For each of the two reference tables (the first based on the three SO scenarios and the second based on the three UO scenarios), we performed ABC analyses on 500 pods per scenario and per prior distribution set (total of 12 000 pods analysed). For each pod, we estimated the posterior probabilities of each of the three competing scenario by polychotomous logistic regression (Cornuet *et al.* 2008) on the 1% of data sets of the reference table closest to the pod. The selected scenario was that with the highest posterior probability value.

It should be stressed that the two competing scenario sets (SO and UO reference tables) are qualitatively equivalent, differing only in terms of the direct use (SO) or nonuse (UO) of the native samples as sources. Thus, when a pod is simulated according to a scenario absent from the reference table, we still have an expected result. For example, if the pod is simulated according to scenario SB (the invasive population originates from the sampled cluster B of the native area; Fig. 2) and the reference table is the UO table (which includes scenarios with unsampled origin UA, UB and UAB; Fig. 2), we would still expect the scenario selected to be UB, as both scenarios SB and UB indicate introduction from a population belonging to the native cluster B. This made it possible to determine which of the competing scenario sets, between SO and UO, was the most prone to error, in choices between introduction from cluster A, cluster B and an admixture of clusters A and B.

ABC analyses on real data sets

Origin of biocontrol strains. In a first set of ABC analyses, we independently considered each of the six biocontrol samples (five European and one American), using the two native population clusters inferred from the population structure analyses and an admixture between them (see Results section) as potential source populations. For each of the six analyses (one for each biocontrol sample) and for each inferred native cluster, we used the native samples displaying the lowest mean pairwise F_{ST} with nonnative populations (i.e. biocontrol and invasive populations; see Results section). The use of other native population samples did not change our conclusions (results not shown). Each ABC analysis was carried out twice: once with an SO scenario design and once with a UO scenario design. Parameter priors were

Table 1 Two sets of parameter distributions for the demographic, historical and mutation parameters used in ABC analyses on controlled simulated data sets

	Broad parameter distribution set						HA-like parameter distribution set					
Parameters	Distribution or value	Mean	Median	Mode	Quantile 2.5%	Quantile 97.5%	Distribution or value	Mean	Median	Mode	Quantile 2.5%	Quantile 97.5%
NS_A	Uniform [100–20 000]	10 023	10 010	NA	590	19 520	10 000	NA	NA	NA	NA	NA
NS_B	Uniform [100–20 000]	10 023	10 010	NA	590	19 520	15 000	NA	NA	NA	NA	NA
NS_{anc}	Uniform [100–20 000]	10 023	10 010	NA	590	19 520	10 000	NA	NA	NA	NA	NA
NS_{inv}	Uniform [100–20 000]	10 023	10 010	NA	590	19 520	Uniform [10 000–15 000]	12 520	12 521	NA	10 121	14 884
NF_{inv}	Loguniform [2–1000]	162	45	2	2	862	Loguniform [2–200]	41	17	2	2	178
BD_{inv}	Uniform [0–5]	2.5	2.5	NA	0	5	Uniform [1–5]	3	3	NA	1	5
t_{anc}	Uniform [100–3000]	1858	1940	NA	380	2,960	800	NA	NA	NA	NA	NA
t_{inv}	50	NA	NA	NA	NA	NA	50	NA	NA	NA	NA	NA
t_{uA} and t_{uB}	Loguniform [50–3000]	475	260	50	50	1980	Loguniform [50–800]	265	190	50	50	750
ar	Uniform [0.1–0.9]	0.5	0.5	NA	0.12	0.88	Uniform [0.1–0.9]	0.5	0.5	NA	0.12	0.88
Mean μ	Uniform $[10^{-5}–10^{-3}]$	5.0×10^{-4}	5.0×10^{-4}	NA	3.5×10^{-5}	9.8×10^{-4}	5.10^{-5}	NA	NA	NA	NA	NA
Mean *P*	Uniform [0.1–0.3]	0.2	0.2	NA	0.10	0.29	0.22	NA	NA	NA	NA	NA
Mean μSNI	Uniform $[10^{-8}–10^{-4}]$	5.0×10^{-5}	5.0×10^{-5}	NA	2.5×10^{-6}	9.7×10^{-5}	2.10^{-5}	NA	NA	NA	NA	NA

Historical and demographic parameters are detailed in Fig. 2. The microsatellite loci were assumed to follow a generalized stepwise mutation model (Estoup *et al.* 2002) with three parameters (Pascual *et al.* 2007; Verdu *et al.* 2009; Lombaert *et al.* 2010): the mean mutation rate (mean μ), the mean parameter of the geometric distribution (mean *P*) of length in terms of the number of repeats of mutation events and the mean mutation rate for single nucleotide instability (mean μSNI). Each locus has a possible range of 40 contiguous allelic states and is characterized by individual μ_{loc} drawn from a gamma (mean = mean μ and shape = 2), P_{loc} drawn from a gamma (mean = mean *P* and shape = 2) and μSNI_{loc} drawn from a gamma (mean = mean μSNI and shape = 2) distribution. Note that for a loguniform[*x*;*y*] distribution, log(*x*) and log(*y*) are the bounds of a uniform distribution. The two sets of parameters were used either to simulate a reference table (broad parameter distribution set) or to simulate pseudo-observed genetic data sets (broad parameter distribution set or HA-like parameter distribution set). The broad parameter distribution set aimed at considering a large set of possible levels of genetic structure and diversity. The HA-like parameter distribution set aimed at mimicking the low level of differentiation, and the high level of diversity found within the native area of *Harmonia axyridis* (see Fig. S1, Supporting information).

identical to those for the 'broad parameter distribution set' used in the simulation analyses (Table 1), assuming 2.5 generations per year for historical parameters and with a few exceptions because of the particular nature of biocontrol populations, which differ from invasive populations: biocontrol populations were assumed to maintain a low effective size remaining constant over time since their collection (i.e. NS_{inv} = log uniform distribution [10–1000]). The steps of the ABC were as described in the previous section.

Worldwide routes of invasion. As suggested by our ABC analyses on controlled simulated data sets (see Results section), geographical genetic structure in the native area of *H. axyridis* may have had an impact on the worldwide invasion routes inferred by Lombaert *et al.* (2010). We therefore performed ABC treatments on the worldwide *H. axyridis* data set, taking into account the possibility that the structured native area may have been incompletely sampled (see the Results section for more details). Because ABC methods for scenario comparison provide relative posterior probabilities with no information on the goodness of fit, we then used the model checking option of DIYABC 1.0 on the final worldwide invasion scenario inferred (as described by Cornuet *et al.* 2010) to determine whether this scenario matches well with the observed genetic data for *H. axyridis*. Briefly, if a model (here, an invasion scenario) fits the observed data correctly, then data simulated under this model with parameters drawn from their posterior distribution should be close to the observed data (Gelman *et al.* 1995) (pp. 159–163). The lack of fit of the model to the data with respect to the posterior predictive distribution can be measured by determining the frequency at which the observed summary statistics are extreme with respect to the test statistic (here, our simulated summary statistics) distribution (hence defining a tail-area probability or *P*-value, for each summary statistic). We simulated 2×10^6 data sets under the final *H. axyridis* invasion scenarios inferred in this study. We then obtained a 'posterior sample' of 2×10^4 values of the posterior distributions of parameters through a rejection step based on Euclidian distances and a linear regression post-treatment (Beaumont *et al.* 2002). We simulated 10^4 data sets with parameter values drawn, with replacement from this 'posterior sample'. Our set of test statistics included the summary statistics used for ABC analysis and two previously unused statistics: the shared allele distances (Chakraborty & Jin 1993) and $(\delta\mu)^2$ distances (Goldstein *et al.* 1995) between each population pair. We did this to reduce the conservative bias associated with the use of summary statistics previously selected for ABC analysis as test statistics (Cornuet *et al.* 2010). Each observed test statistic was compared with 10^4 simulated test statistics, and its p-value was calculated.

Results

Genetic variation within populations

We genotyped a total of 271 individuals originating from nine sites sampled within the native range (Table S1, Supporting information). The level of polymorphism estimated over all native sites was substantial, with a mean number of alleles per locus of 12.9. Allelic richness at 18 microsatellite loci, corrected for 20 individuals per sample (AR_{18}), ranged from 5.26 alleles per locus for the Kazakhstan sample (N-Kazak) to 6.59 for one of the Japanese samples (N-Japan1) (Fig. S2, Supporting information). The two European biocontrol samples for which AR_{18} could be calculated (EB-INRA06 and EB-Biotop) displayed much lower levels of diversity, with <2.4 alleles per locus. Other European biocontrol populations also displayed substantially lower diversities: allelic richness at 10 microsatellite loci corrected for 13 individuals per sample (AR_{10}) was at least 30% lower than that for the least diverse native sample. By contrast, the American biocontrol sample had an AR_{10} very similar to that of native populations. All invasive populations displayed high genetic diversities (Fig. S2, Supporting information). However, slightly lower diversity values were obtained for the African population (I-AF), and markedly lower diversity values were obtained for the South American population (I-SA).

Genetic variation between populations

Most pairwise comparisons between populations collected within the native area showed significant genotypic differentiation (Table S2, Supporting information). However, despite the large geographical distances between our sample sites (mean spatial distance = 2700 km), pairwise F_{ST} estimates were low, with a mean of 0.013 and values ranging from −0.006 to 0.035 (Table S2, Supporting information). By contrast, the level of genetic differentiation between European biocontrol samples was high, with a mean F_{ST} of 0.231. European biocontrol samples systematically yielded their lowest F_{ST} values with the Yunnan Chinese sample (N-China2) in the native range (mean F_{ST} between EB samples and N-China2 = 0.206). The American biocontrol sample had low F_{ST} values with native samples, the lowest being 0.017 with the Jilin Chinese sample (N-China3). Genetic differentiation within the invasive range was moderate (mean F_{ST} = 0.064), and the lowest F_{ST} values with populations from the native range were those for the N-China2 or N-China3 sample.

Native samples grouped together in the NJ tree (Fig. 3), with two subclusters, one including the three western samples (N-Russia1, N-Russia2 and N-Kaza) and the other the six eastern samples (N-China1, N-China2, N-China3, N-Japan1, N-Japan2 and N-Korea). Despite the long branches, all European biocontrol samples grouped together, tending to confirm a common origin of these samples.

STRUCTURE analyses (Pritchard *et al.* 2000) of *H. axyridis* individuals sampled within the native area provided consistent results over the 20 runs tested for each *K*. The natural logarithm of the likelihood of the data $\ln(P(X|K))$ increased from $K = 1$ to $K = 2$, for which it was maximal (Fig. S3, Supporting information). The proportion of ancestry from each of the two clusters of each native sample defined two geographical areas identical to those suggested by the NJ tree: a 'western cluster' and an 'eastern cluster' (Fig. 1). The use of other STRUCTURE models [with or without (i) admixture, (ii) correlated allele frequencies or (iii) sampling location information] gave similar results. $K = 1$ had the highest likelihood in a few cases, but this is not surprising given the low level of differentiation between populations. BAPS spatial clustering analysis (Corander *et al.* 2004) confirmed the existence of these two geographical clusters (Fig. 1). Mean pairwise F_{ST} was 0.000 and 0.007 within the western and eastern clusters, respectively, whereas the mean F_{ST} between populations from different clusters was 0.021 (Table S2, Supporting information). The level of differentiation within clusters was thus low, but still significant for many pairwise comparisons (Table S2, Supporting information), revealing slight structuring of the populations within the western and eastern clusters.

A significant correlation between the measures of genetic differentiation and geographical distance was found within the native area ($r^2 = 0.304$; $P < 10^{-2}$; slope = 0.008). However, this correlation most probably reflected the presence of two populational groups separated by large geographical distances rather than a continuous pattern of isolation by distance. In agreement with this, we did not found any significant correlation when considering only samples from Eastern Asia despite large geographical distances between the sampled sites (six samples: $r^2 = 0.009$; $P = 0.534$).

ABC in a structured native range: simulation-based study

We first considered the pseudo-observed data sets (pods) simulated with the 'broad parameter distribution set' (Table 1). When the true scenarios were those for

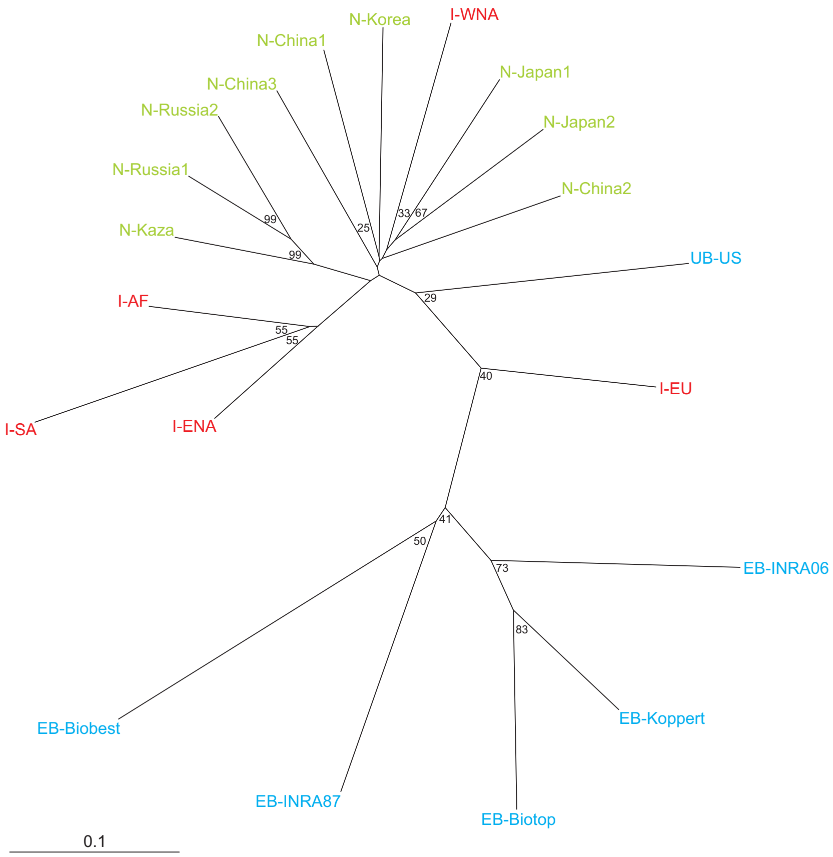

Fig. 3 Neighbour-joining tree for *Harmonia axyridis* population samples based on the chord distance of Cavalli-Sforza & Edwards (1967). Population code names are as in Table S1 (Supporting information). Native population samples are shown in green, invasive population samples in red and biocontrol samples in blue. Bootstrap values calculated over 1000 replications are given as percentages (only values >20% are shown).

which the actual source population of the invasive population had been sampled (SA, SB or SAB), the proportion of error (i.e. inference of an incorrect population cluster as the source) was very low, whatever the reference table used (SO or UO reference table, Table 2). By contrast, when the true scenarios were those in which the actual source of the invasive population had not been sampled (but a genetically different population from the same cluster, i.e. UA, UB or UAB), large error rates were obtained when using the reference table assuming that the source population has been sampled (the SO reference table; Table 2). These errors corresponded principally to incorrect selection of the admixed scenario SAB in 42.4% of cases when the true scenario was a single nonadmixed introduction (scenario UA or scenario UB). Error rates were markedly lower if it was assumed that the actual source population had not been sampled (the UO reference table; Table 2). In particular, the frequency of incorrect selection of the admixture scenario decreased to 8.4%.

We then considered the pods simulated with the 'HA-like parameter distribution set', chosen because they fitted the real *H. axyridis* situation more closely (Table S2 and Fig. S1, Supporting information). As for the 'broad parameter distribution set', results were generally better when the reference table was simulated assuming that the actual source had not been sampled (UO reference table; Table 2). Overall, error rates shown in Table 2 were remarkably high, especially when considering the 'HA-like parameter distribution set'. Most errors actually corresponded to pods obtained with *ar* (admixture rates) values close to the upper and lower limits of this parameter distribution (i.e. close to 0.1 or 0.9) and/or with small t_{anc} (splitting time between the two native population clusters from an ancestral population) values (see Table S3, Supporting information for type I and type II error rates when *ar* is intermediate and t_{anc} is large).

In conclusion, the UO reference table gave globally better inferences about invasion pathways, generating lower type I and type II error rates than the SO reference table (Table 2). In particular, we found that use of the UO reference table substantially reduced the risk of finding admixture between native source population clusters when there was none and only slightly increased the risk of selecting simple scenarios without admixture when there was admixture.

Origin of biocontrol strains

The native population samples displaying the lowest pairwise F_{ST} with all nonnative populations were the N-Kaza sample from the western cluster (mean $F_{ST} = 0.139$) and the N-China2 sample from the eastern cluster (mean $F_{ST} = 0.114$). We thus used these two native samples as representative of the western and

Table 2 Confidence in scenario selection based on ABC analyses on pseudo-observed data sets

		Competing scenario set (reference table)			
		Sampled origin (SO)		Unsampled origin (UO)	
Pods' parameter distribution set	Scenario considered	Type I error	Type II error	Type I error	Type II error
Broad	SA	0.032	0.052	0.024	0.052
	SB	0.024	0.036	0.016	0.060
	SAB	0.176	0.028	0.224	0.020
	S mean	0.077	0.039	0.088	0.044
	UA	0.440	0.056	0.096	0.100
	UB	0.424	0.040	0.096	0.064
	UAB	0.176	0.424	0.304	0.084
	U mean	0.347	0.173	0.165	0.083
HA-like	SA	0.120	0.146	0.024	0.196
	SB	0.140	0.128	0.056	0.164
	SAB	0.520	0.116	0.688	0.024
	S mean	0.260	0.130	0.256	0.128
	UA	0.392	0.112	0.208	0.200
	UB	0.320	0.092	0.192	0.200
	UAB	0.376	0.340	0.616	0.108
	U mean	0.363	0.181	0.339	0.169

The compared scenarios are detailed in Fig. 2, and parameter distributions are given in Table 1. Type I error: proportion of cases in which the scenario considered is excluded but is actually the true one. Type II error: proportion of cases in which the scenario considered is selected but is not the true one.

eastern native clusters, respectively, in all ABC analyses of real data sets. Using other samples gave qualitatively similar results (data not shown).

With the UO reference table, all ABC analyses performed on the separate biocontrol strains gave the highest posterior probability for the eastern native cluster being the origin (Table S4, Supporting information). Interestingly, when the SO reference table was used with the EB-INRA87 sample, the confidence interval for the probability of an eastern native cluster origin almost entirely overlapped with that for the admixed scenario. This made it impossible to distinguish between these two scenarios and highlights the advantages of using the UO reference table, as previously demonstrated in the simulations. Taking into account the inferred eastern native cluster origin of all biocontrol strains, we then showed, by ABC, that all European biocontrol strains were actually derived from the same ancestral population (see Table S5, Supporting information). This result confirmed that the main biofactories in Europe had been rearing *H. axyridis* samples originating from the same population collected by INRA in the eastern part of the native area in 1982.

Worldwide routes of invasion

As described by Lombaert *et al.* (2010), we performed five serial nested ABC analyses of invasion scenarios involving successive *H. axyridis* outbreaks (eastern North America, western North America, Europe, South America and then South Africa). Each analysis was thus carried out by simulating a new reference table taking into account the previous result. For example, the most likely origin of the ENA outbreak inferred in the first ABC analysis was included in the second analysis when this population became a potential source of the western North American outbreak. As for the parameters of the scenarios, the same prior were used at every steps (i.e. the posterior distributions of parameters from an analysis were not used as prior in the next analysis). Samples, priors and scenarios were as described by Lombaert *et al.* (2010), with a few exceptions: (i) we used the western and eastern native clusters as potential sources (with N-Kaza and N-China2 as representative samples); (ii) the competing scenarios involving a native sample were drawn from the UO scenario set design, that is taking into account the possibility that the actual native source population might not have been sampled and (iii) we added the American biocontrol sample (UB-US sample) as a potential source for the eastern and western North American outbreaks only. As reported previously, all biocontrol populations used in the analyses were derived from the eastern native cluster. Information about the set of scenarios considered and prior distributions are given in Table 3 and Table S6 (Supporting information), respectively. The worldwide routes of invasion of *H. axyridis* inferred by Lombaert *et al.* (2010) were confirmed by this new ABC analysis (Table 3). The main new findings were that the ENA outbreak was the result of an admixture between the eastern and the western native clusters (the use of other native population samples did not change our conclusions; Table S7, Supporting information), with each cluster making an approximately equal contribution (admixture rate estimated at 57% for the eastern native cluster, 95% CI: [16%–86%]). By contrast, the western North American outbreak originated exclusively from the eastern native cluster. The relationships of the samples in the NJ tree analysis were consistent with our ABC-based conclusions (Fig. 3).

To better evaluate to what extend the admixed native origin of ENA (referred to hereafter as ‘scenario5’ of analysis 1) could be trusted, we computed the type I and type II errors of this scenario in analysis 1. To do so, we simulated 100 pods per scenario. As expected from our previous simulation study, type I error rate was substantial with a value of 0.45. More importantly, however, type II errors were very low: the mean value was equal to 0.02 with values ranging from 0 to 0.07. It is worth pointing out that type I and type II errors do not take into account the posterior probability that was actually found with the real dataset. Following Fagundes *et al.* (2007), we used our estimations of type I and type II error rates to compute the probability that scenario5 was the correct scenario given our observation that $P_{\text{scenario5}} = 0.6242$ as $\Pr(P_{\text{scenario5} \geq 0.6242} \mid \text{scenario5 is true}) / \sum_{i=1}^{10} \Pr(P_{\text{scenario5} \geq 0.6242} \mid \text{scenario}i \text{ is true}) = 0.8649$. These results reinforce the overall conclusion of our study, specially the admixed origin of ENA.

Model checking was carried out for the final selected worldwide invasion scenario that includes the five *H. axyridis* invasive outbreaks (see Fig. S4, Supporting information). We found that the observed values of only six summary statistics (none of those not used for ABC inferences) of a total of 279 (i.e. 2.2%) lay in the tail of the probability distribution of statistics calculated from the posterior simulation (i.e. $P < 0.05$ or $P > 0.95$). Because this analysis may suffer from nonindependence between the summary statistics, we also performed a principal component analysis (PCA). Figure S5 (Supporting information) illustrates the result of a PCA in the space of the summary statistics. It shows that PCA points simulated from the posterior predictive distribution nicely grouped and relatively well centred on the target point corresponding to the real data set. Altogether, these results indicate that the final selected worldwide invasion scenario provides a satisfying description of our real *H. axyridis* data set.

Table 3 Description of the five ABC analyses attempting to retrace step by step the worldwide invasion routes of *Harmonia axyridis* and posterior probabilities of the selected (most likely) scenarios in each ABC analysis

Analysis and target outbreak	Potential source populations (+admixture between all source pairs)	Number of scenarios	Selected scenario	Posterior probability of selected scenario

Analysis 1: Eastern North America	(1) Western native cluster;
	(2) Eastern native cluster;
	(3) American biocontrol;
	(4) European biocontrol
		10	Admixture: western native cluster + eastern native cluster	0.6242 [0.5767,0.6717]

Analysis 2: Western North America	(1) Western native cluster;
	(2) Eastern native cluster;
	(3) American biocontrol;
	(4) European biocontrol;
	(5) Eastern North America
		15	Eastern native cluster	0.4425 [0.3746,0.5105]

Analysis 3: Europe	(1) Western native cluster;
	(2) Eastern native cluster;
	(3) European biocontrol;
	(4) Eastern North America;
	(5) Western North America
		15	Admixture: European biocontrol + eastern North America	0.8134 [0.7107,0.9160]

Analysis 4: South America	(1) Western native cluster;
	(2) Eastern native cluster;
	(3) European biocontrol;
	(4) Eastern North America;
	(5) Western North America
		15	Eastern North America	0.9489 [0.9315,0.9663]

Analysis 5: Africa	(1) Western native cluster;
	(2) Eastern native cluster;
	(3) European biocontrol;
	(4) Eastern North America;
	(5) Western North America;
	(6) Europe;
	(7) South America
		28	Eastern North America	0.8692 [0.7422,0.9961]

In analyses 1–4, posterior probabilities were calculated by polychotomous logistic regression on the simulations corresponding to the 1% smallest Euclidean distances. In analysis 5, because of computational issues owing to the large number of scenarios compared and summary statistics, posterior probabilities were calculated in two steps: (i) we calculated posterior probabilities on the 0.01% smallest Euclidean distances by the direct approach (Cornuet *et al.* 2008) and then removed 11 of the 28 compared scenarios for which the direct posterior probability was lower than 10^{-3} and (ii) we re-estimated the posterior probabilities of each of the 19 remaining competing scenarios by polychotomous logistic regression on the 0.5% smallest Euclidean distances. 95% confidence intervals (CI) are shown in square brackets. The 95% CI of the selected scenarios never overlapped with those of competing scenarios. The samples used in the analyses were as follows (see Table S1, Supporting information): western native cluster = N-Kazak; eastern native cluster = N-China2; American biocontrol = UB-US; European biocontrol = EB-INRA87; eastern North America = I-ENA; western North America = I-WNA; Europe = I-EU; South America = I-SA; Africa = I-AF.

Discussion

Sampling effort, genetic structure within the native range and false admixture

Comprehensive sampling of a species distribution area is often impossible for practical reasons (e.g. difficulties reaching some locations and/or poor knowledge of the exact range). This is the case for *H. axyridis*, which has a large and imperfectly known native area (e.g. Poutsma *et al.* 2008). Sampling effort and design are recurrent issues in population genetics, and several studies have shown that incomplete sampling may introduce bias into inferences relating to genetic structure and connectivity between populations (Waples & Gaggiotti 2006; Muirhead *et al.* 2008). Bayesian clustering methods have been improved to incorporate sampling scheme or space in models (Corander *et al.* 2004; Guillot *et al.* 2005; Hubisz *et al.* 2009), but they still cannot fully compensate for the absence of samples from a number of locations in the native range, for genetically structured populations.

Our results based on the analyses of controlled simulated data sets show that when an invasive species is genetically structured in its native area, the ability of ABC analyses to infer invasion routes correctly may be jeopardized by incomplete sampling of the native area. In particular, in the simplest case when genetic structure exists within each of two main genetic clusters (as found for *H. axyridis*), ABC analyses often erroneously select a scenario of admixture between the two clusters when the true scenario is a simple origin, without admixture, from an unsampled population from one of the two clusters. Fortunately, an ABC package, such as DIYABC, can incorporate the possibility that the native samples used in the analysed data set are not the direct source populations, by modelling unsampled populations genetically differentiated from those samples to some extent. This approach led to a halving of type I and type II errors with the broader parameter distribution set used in our analyses. This was because of a large decrease in the frequency of erroneous selection of admixture scenarios. In addition, it is worth stressing that the simulation of unsampled populations may also make it easier to deal with too large numbers of slightly differentiated samples, which would make the already cumbersome ABC analyses impossible if they were all used as potential source populations.

The robust identification of admixture between two or more native population clusters as the origin of an invasive outbreak is crucial in the field of invasion biology. Admixture can produce new recombinant genotypes and compensate for the loss of diversity and additive genetic variance potentially following founder events. Admixture has therefore been identified as one of the key factors underlying invasion success, through its effects on the process of adaptation following establishment (Wares *et al.* 2005; Facon *et al.* 2006; Keller & Taylor 2008). It is therefore important to include admixture events between native potential sources as competing invasion route scenarios. This is particularly true given that classical population genetic statistics usually provide little information about this phenomenon and may be misleading in some cases (e.g. Lombaert *et al.* 2010). However, it is also essential to avoid the selection of false admixture scenarios in ABC analysis, to prevent erroneous interpretations of the evolutionary factors instrumental to the success of an invasion.

Genetic structure within the native range of H. axyridis

Our genetic analyses inferred a clear genetic structure of *H. axyridis* in its native area, consisting of two distinct geographical clusters with (i) Kazakhstan and central Siberia in the west and (ii) China, Korea and Japan in the east. Consistent with this pattern, an analysis of phenotypic traits, such as elytral patterns, indicated that *H. axyridis* could be divided into two geographical groups, with a dividing line between them located in the zone of the Baikal fracture (Dobzhansky 1933; Blekhman 2008; Blekhman *et al.* 2010). The observed genetic structure could be due to the occurrence of a natural barrier, such as the dry central Asia plateau and the Baikal rift zone, which may limit gene flow between the two parts of the native area of *H. axyridis*. Furthermore, as suggested by Blekhman *et al.* (2010), natural populations of *H. axyridis* may have split into two separate geographical groups during the last Pleistocene glaciation, subsequently merging during the Holocene warming, leading to hybridization around the Baikal fracture. Bayesian clustering methods such as STRUCTURE (Pritchard *et al.* 2000) tends to overestimate genetic structure when analysing a data set characterized by genetic isolation by geographical distances (IBD, e.g. Frantz *et al.* 2009). In the case of *H. axyridis*, however, the absence of significant correlation between genetic and geographical distances within the eastern cluster (where six of nine samples were collected) suggests that the significant correlation that was found considering all nine native samples most probably reflected the presence of two populational groups separated by large geographical distances rather than a continuous pattern of isolation by distance. In agreement with this, no cline was found on morphological traits within both groups despite strong differences between groups (Blekhman *et al.* 2010). Additional genetic data, particularly for samples collected from the intermediate area between the Russian administrative regions of Irkutsk and Amur and Mongolia, are required to shed light on the evolutionary factors involved in the genetic structure of *H. axyridis* in its native range.

As predicted, the ABC analyses performed to elucidate the native origin of the *H. axyridis* biocontrol samples were enhanced by the simulation of unsampled native populations. We found that all biocontrol samples originated from the eastern cluster of the native area of *H. axyridis*, and the validity of this result was further supported by subsequent ABC analysis, which confirmed, with a high posterior probability, that all of our European biocontrol samples were derived from a single ancestral population sampled from the native area of *H. axyridis* by INRA in 1982. This finding was also supported by the monophyletic relationship of these samples in the NJ tree. Finally, the inferred eastern origin of the biocontrol samples analysed here is consistent with the available historical information: the original European biocontrol population was sampled in China (Beijing, Ongagna *et al.* 1993) and the American biocontrol sample used in this study originated from the far east of Russia (Ussuriysk), according to the USDA database (http://www.ars-grin.gov/cgi-bin/nigrp/robo/f941s.pl?50902).

Worldwide invasion routes of H. axyridis*: what's new?*

The overall history of *H. axyridis* introduction inferred by Lombaert *et al.* (2010) was largely supported by these ABC analyses. We confirmed that the recent burst of worldwide invasion by *H. axyridis* has followed a bridgehead scenario, in which an invasive population in eastern North America acted as the source of the colonists invading the European, South American and African continents, with some admixture with a biocontrol strain in Europe. The two North American outbreaks were confirmed to have originated independently from the native area. The single American biocontrol sample included in our ABC analyses was not involved in any of the American outbreaks. However, although an accidental origin has been suggested before (Koch 2003), many different native populations have been imported and used for biocontrol purposes in North America, so a biocontrol origin cannot be excluded.

Posterior model checking for the final worldwide scenario of *H. axyridis* invasion gave good results. This suggests that the simulation of an incompletely sampled, but structured, native area in the analysis of *H. axyridis* invasion routes provides a good fit with the real dataset. In addition, these ABC analyses made it possible to make further inferences about the origin of the North American invasive populations. The source of the western North American (WNA) outbreak was the eastern cluster of native area of *H. axyridis*, whereas the ENA outbreak resulted from an admixture between the two native clusters, with each cluster making an approximately equal genetic contribution. This admixed origin of the ENA outbreak is of particular interest. First, this result was, to some extent, unexpected, given the known history of *H. axyridis* biocontrol (Tedders & Schaefer 1994; Krafsur *et al.* 1997) and current airline transportation networks (e.g. Tatem & Hay 2007), both of which identified eastern Asia as the most likely origin of the American outbreaks, as confirmed for the WNA outbreak. Second, the ENA population has served as a bridgehead for worldwide invasion by *H. axyridis*, and the finding that it is probably a genetically admixed population has important implications for our understanding of the key factors involved in the invasion success of this ladybird. Indeed, after decades of unsuccessful acclimation of biocontrol strains, genetic admixture in the ENA population may have facilitated adaptation by allowing the appearance of new gene combinations. However, it remains unknown whether admixture occurred before or after the introduction. The sampling and genotyping of populations from the contact zone between the two native clusters might provide us with some answers to this question. Finally, Facon *et al.* (2011) recently found that deleterious mutations at life history traits important for invasion success have been purged in the ENA bridgehead population, probably due to bottleneck event(s) of appropriate intensity. Additional studies are required to assess the relative and/or complementary roles of admixture, bottlenecks and purging in the success of this key *H. axyridis* outbreak.

Acknowledgements

We thank the SDAR service of Sophia-Antipolis, Nicole Bareste, Paulette Flacchi, Aurélie Blin and Alexandra Auguste for technical and administrative assistance. We also thank our colleagues T. Malausa, C. Borges, R. Koch, J. Lundgren, M. Majerus and R. Staals for additional *H. axyridis* samples. This work was funded by the French ANR Biodiversité #ANR-06-BDIV-008-01 Grant.

References

Adriaens T, Branquart E, Maes D (2003) The multicoloured Asian ladybird *Harmonia axyridis* Pallas (Coleoptera : Coccinellidae), a threat for native aphid predators in Belgium? *Belgian Journal of Zoology*, **133**, 195–196.

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

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300.

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.

Blekhman AV (2008) Population variation of elytral ridge occurrence in ladybirds *Harmonia axyridis* Pallas. *Russian Journal of Genetics*, **44**, 1351–1354.

Blekhman AV, Goryacheva II , Zakharov IA (2010) Differentiation of *Harmonia axyridis* Pall. according to polymorphic morphological traits and variability of the mitochondrial COI gene. *Moscow University Biological Sciences Bulletin*, **65**, 174–176.

Blum MGB, Francois O (2010) Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, **20**, 63–73.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics*, **19**, 233–257.

Chakraborty R, Jin L (1993) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. *EXS*, **67**, 153–175.

Chapin J, Brou V (1991) *Harmonia axyridis* (Pallas), the third species of the genus to be found in the United States (Coleoptera: Coccinellidae). *Proceedings of the Entomological Society of Washington*, **93**, 630–635.

Ciosi M, Miller NJ, Kim KS *et al.* (2008) Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various reductions of genetic diversity. *Molecular Ecology*, **17**, 3614–3627.

Corander J, Waldmann P, Marttinen P, Sillanpaa MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Cornuet JM, Santos F, Beaumont MA *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Cornuet JM, Ravigne V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**, Art. No 401.

Csillery K, Blum MGB, Gaggiotti OE, Francois O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, **25**, 410–418.

Dobzhansky T (1933) Geographical variation in lady-beetles. *The American Naturalist*, **67**, 97–126.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, **19**, 4113–4130.

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.

Facon B, Genton BJ, Shykoff J *et al.* (2006) A general eco-evolutionary framework for understanding bioinvasions. *Trends in Ecology & Evolution*, **21**, 130–135.

Facon B, Hufbauer RA, Tayeh A *et al.* (2011) Inbreeding depression is purged in the invasive insect *Harmonia axyridis*. *Current Biology*, **21**, 424–427.

Fagundes NJR, Ray N, Beaumont MA *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.

Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

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

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis*. Chapman and Hall, New York.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**, 463–471.

Goudet J (2002) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3.2). Updated from Goudet (1995). Available from http://www2.unil.ch/popgen/softwares/fstat.htm.

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.

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Hamilton G, Currat M, Ray N *et al.* (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*, **170**, 409–417.

Hardy OJ, Vekemans X (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, Art. No 26.

Keller SR, Taylor DR (2008) History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. *Ecology Letters*, **11**, 852–866.

Koch RL (2003) The multicolored Asian lady beetle, *Harmonia axyridis*: a review of its biology, uses in biological control, and non-target impacts. *Journal of Insect Science*, **3**, 1–16.

Kolbe JJ, Glor RE, Schettino LRG *et al.* (2004) Genetic variation increases during biological invasion by a Cuban lizard. *Nature*, **431**, 177–181.

Krafsur ES, Kring TJ, Miller JC *et al.* (1997) Gene flow in the exotic colonizing ladybeetle *Harmonia axyridis* in North America. *Biological Control*, **8**, 207–214.

LaMana ML, Miller JC (1996) Field observations on *Harmonia axyridis* Pallas (Coleoptera: Coccinellidae) in Oregon. *Biological Control*, **6**, 232–237.

Leberg PL (2002) Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology*, **11**, 2445–2449.

Loiseau A, Malausa T, Lombaert E, Martin JF, Estoup A (2009) Isolation and characterization of microsatellites in the harlequin ladybird, *Harmonia axyridis* (Coleoptera, Coccinellidae), and cross-species amplification within the family Coccinellidae. *Molecular Ecology Resources*, **9**, 934–937.

Lombaert E, Guillemaud T, Cornuet JM *et al.* (2010) Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird. *PLoS ONE*, **5**, e9743.

Miller N, Estoup A, Toepfer S *et al.* (2005) Multiple transatlantic introductions of the western corn rootworm. *Science*, **310**, 992.

Muirhead JR, Gray DK, Kelly DW *et al.* (2008) Identifying the source of species invasions: sampling intensity vs. genetic diversity. *Molecular Ecology*, **17**, 1020–1035.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nunes MA, Balding DJ (2010) On optimal selection of summary statistics for approximate Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, **9**, Art. No 34.

Ongagna P, Giuge L, Iperti G, Ferran A (1993) Life-cycle of *Harmonia axyridis* (Col, Coccinellidae) in its area of introduction—South-Eastern France. *Entomophaga*, **38**, 125–128.

Pascual M, Chapuis MP, Mestres F *et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, **16**, 3069–3083.

Poutsma J, Loomans AJM, Aukema B, Heijerman T (2008) Predicting the potential geographical distribution of the harlequin ladybird, *Harmonia axyridis*, using the CLIMEX model. *BioControl*, **53**, 103–125.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Raymond M, Rousset F (1995a) An exact test for population differentiation. *Evolution*, **49**, 1280–1283.

Raymond M, Rousset F (1995b) Genepop (version. 1.2), a population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Saini E (2004) Presencia de *Harmonia axyridis* (Pallas) (Coleoptera: coccinellidae) en la provincia de Buenos aires. Aspectos biologicos y morfologicos. *RIA*, **33**, 151–160.

Saitou N, Nei M (1987) The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

Stals R, Prinsloo G (2007) Discovery of an alien invasive, predatory insect in South Africa: the multicoloured Asian ladybird beetle, *Harmonia axyridis* (Pallas) (Coleoptera: Coccinellidae). *South African Journal of Science*, **103**, 123–126.

Tatem AJ, Hay SI (2007) Climatic similarity and biological exchange in the worldwide airline transportation network. *Proceedings of the Royal Society B-Biological Sciences*, **274**, 1489–1496.

Tedders WL, Schaefer PW (1994) Release and establishment of *Harmonia Axyridis* (Coleoptera, Coccinellidae) in the Southeastern United States. *Entomological News*, **105**, 228–243.

Verdu P, Austerlitz F, Estoup A *et al.* (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19**, 1–7.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wares JP, Hughes AR, Grosberg RK (2005) Mechanisms that drives evolutionary change: insights from species introductions and invasions. In: *Species Invasions: Insights into Ecology, Evolution and Biogeography* (eds Sax DF, Stachowicz JJ, Gaines SD), pp. 229–257. Sinauer Associates Inc., Sunderland, MA.

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Weir BS, Cockerham C (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

E.L., T.G. and A.E.'s research focuses on the evolutionary biology of invading species, and are currently interested in both the application of methods for historical inferences using molecular markers and the evolution of life history traits in the context of biological invasions. C.E.T. is a PhD student investigating the evolutionary genetics of the *Harmonia axyridis* invasion. L.-J.L.H. is a lecturer in Evolutionary Biology, and is particularly interested in the evolutionary biology of invasions. J.L. is an entomologist, and his research focuses on taxonomy and phylogeny. S.W.'s interest is the ecological manipulation of predacious ladybird application in biological control. H.P.'s research interests are phylogeny and evolution of ladybird beetles (Coleoptera: Coccinellidae) Mass rearing and genetics of *Harmonia axrydis*, *Cryptolaemus montrouzieri*. I.G.'s research interests are in the field of molecular ecology, insects phylogenetics and genetics of insect-bacteria symbiosis. I.A.Z.'s main research interests are population genetics of *Adalia bipunctata* and *Harmonia axyridis* and interaction of insects with symbiotic bacteria. E.J.'s main research interests are phylogeny and evolution of interspecific interactions. R.L.P. is a secondary school teacher (biology) at Clifton College. She has extensive research expertise in intraguild interactions. A.M. is interested in the systematics and biology of mites. He works, among other things, on introduction and invasion pathways of alien spider mites. He is also interested in distribution patterns and modelling of potential distribution of these invasive species. J.L. works since 1970 on behavioural ecology and population dynamics of parasitoids, theoretical and practical aspects of biological control, IPM and sustainable crop production, and anatomy and sensory physiology of parasitoid ovipositors. P.C.'s research group focuses on the integrated management of arthropod pests, with emphasis on the potential of predatory insects and mites for augmentative biological control. His special attention goes out to risk assessment for biological control agents and to the development of rearing systems for predatory arthropods. N.B. is interested in ecology and life history traits of invasive ladybirds. W.J. is currently Director of the USDA, ARS, National Biological Control Laboratory in Mississippi. He has conducted research for over 30 years, focused on biological control of invasive insects and weeds, with special interest in the ecology and behavior of egg parasitoids of pest Heteroptera.

Data accessibility

Sample locations and microsatellite data: DRYAD entry doi:10.5061/dryad.7m0b37bn.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Simulated controlled data sets and corresponding levels of genetic diversity and levels of genetic differentiation between population clusters A and B (continuous circles) and between 'unsampled' populations within each cluster (dashed circles).

Fig. S2 Genetic diversity in the native, biocontrol and invasive population samples of *Harmonia axyridis.*

Fig. S3 Estimated number of population clusters in the native *Harmonia axyridis* samples according to the Bayesian clustering method STRUCTURE.

Fig. S4 Final selected worldwide invasion scenario which includes the five *H. axyridis* invasive outbreaks.

Fig. S5 Graphical representation of the result of a principal component analysis (PCA) in the space of the summary statistics performed on the final selected worldwide invasion scenario.

Table S1 Native, biocontrol and invasive population samples of *Harmonia axyridis* used in this study.

Table S2 Pairwise estimates of F_{ST} between all *Harmonia axyridis* population sample pairs.

Table S3 Confidence in scenario selection based on ABC analyses on pseudo-observed data sets.

Table S4 ABC posterior probabilities of the three competing scenarios modeling the genetic origin of each biocontrol sample within the native area of *Harmonia axyridis* (western native cluster, eastern native cluster or admixture of the western and eastern native clusters).

Table S5 ABC analyses to assess the relationship between the five European Biocontrol populations.

Table S6 Prior distributions of demographic, historic and mutation parameters used in ABC analyses attempting to retrace the worldwide routes of invasion of *Harmonia axyridis*.

Table S7 Inferred origin of the 'Eastern North American' (ENA) outbreak with various combinations of native samples representative of the Western and Eastern native clusters.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

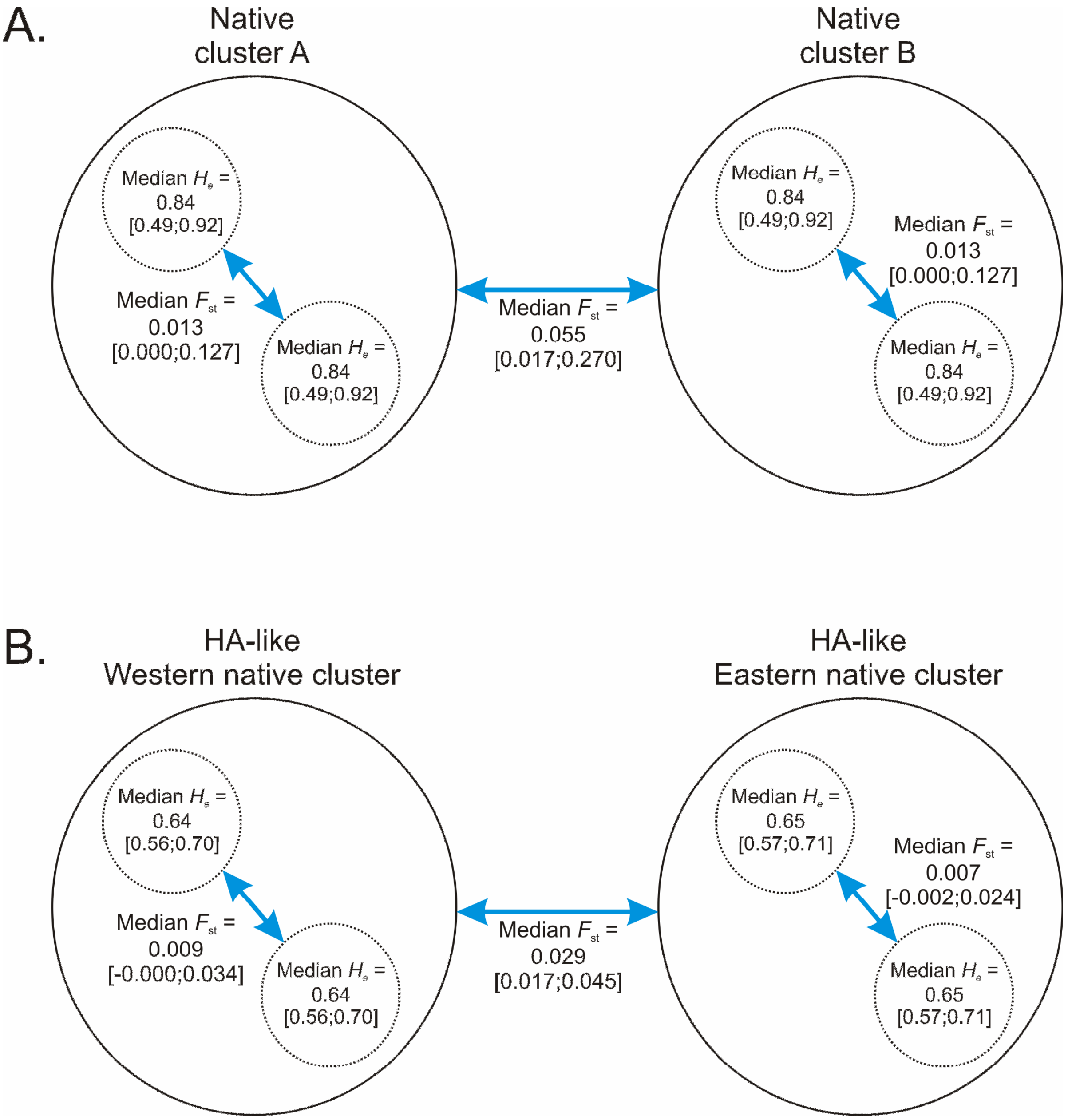
A.
Native cluster A
Native cluster B
Median H_e = 0.84 [0.49;0.92]
Median F_{st} = 0.013 [0.000;0.127]
Median H_e = 0.84 [0.49;0.92]
Median F_{st} = 0.055 [0.017;0.270]
Median H_e = 0.84 [0.49;0.92]
Median F_{st} = 0.013 [0.000;0.127]
Median H_e = 0.84 [0.49;0.92]
B.
HA-like Western native cluster
HA-like Eastern native cluster
Median H_e = 0.64 [0.56;0.70]
Median F_{st} = 0.009 [-0.000;0.034]
Median H_e = 0.64 [0.56;0.70]
Median F_{st} = 0.029 [0.017;0.045]
Median H_e = 0.65 [0.57;0.71]
Median F_{st} = 0.007 [-0.002;0.024]
Median H_e = 0.65 [0.57;0.71]

Figure S1: Simulated controlled data sets and corresponding levels of genetic diversity and levels of genetic differentiation between population clusters A and B (continuous circles) and between "unsampled" populations within each cluster (dashed circles). (A) Pseudo-observed data sets were simulated with the "broad parameter distribution set" (Table 1); (B) pseudo-observed data sets were simulated with the "HA-like parameter distribution set" (Table 1). Median H_e, median F_{ST} and 95% confidence interval (in brackets) were calculated from 10,000 simulated data sets.

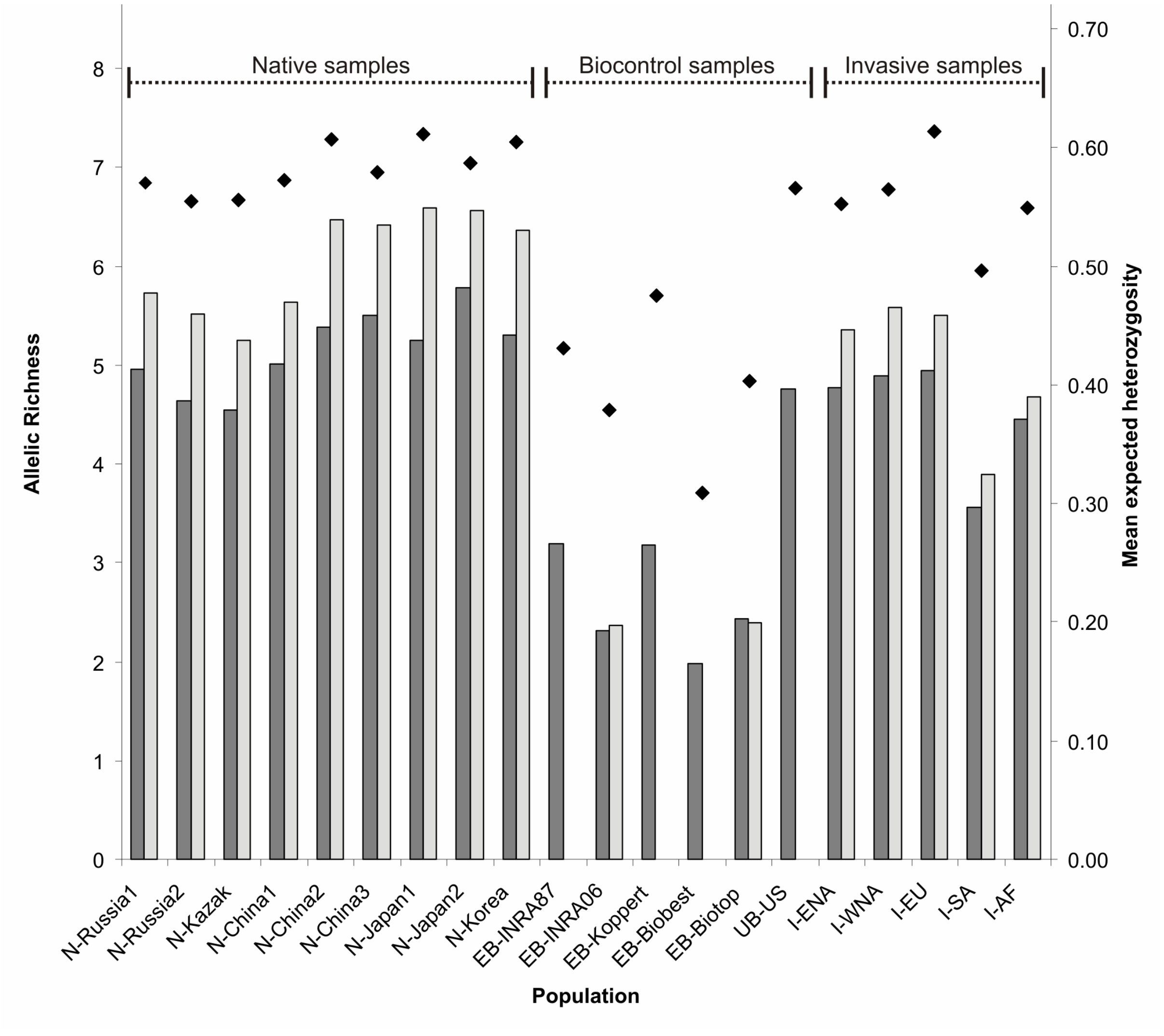

Figure S2: Genetic diversity in the native, biocontrol and invasive population samples of *Harmonia axyridis*.

Several measurements are displayed: expected heterozygosity (*H_e*; black diamond) and average genetic diversity estimated as allelic richness at either 18 microsatellite loci (*AR_{18}*; light gray bars) or a subset of 10 microsatellite loci (*AR_{10}*; dark gray bars). *H_e* are very similar for 10 and 18 loci, thus only the values for 18 loci are shown.

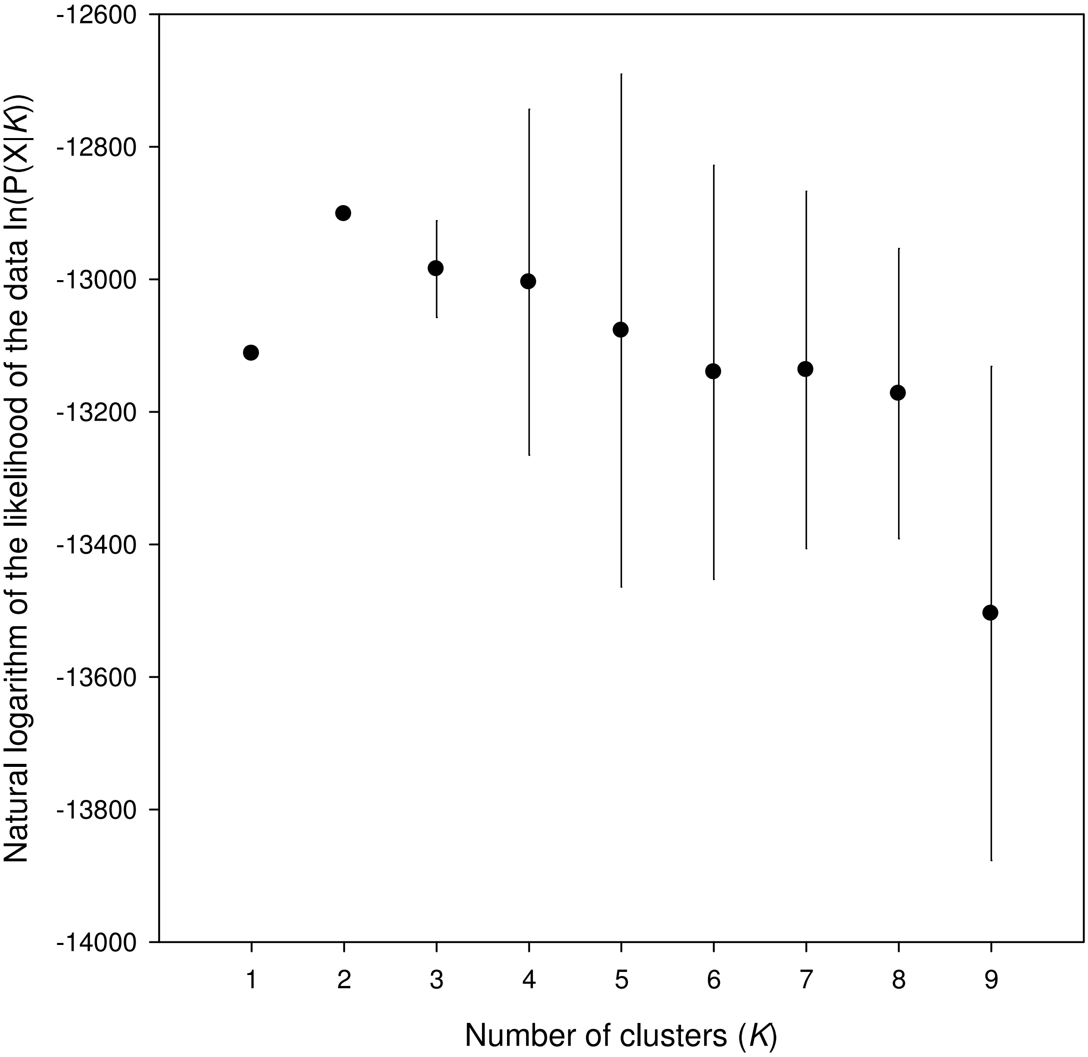

Figure S3: Estimated number of population clusters in the native *Harmonia axyridis* samples according to the Bayesian clustering method STRUCTURE.

The mean (±SD) natural logarithm of the likelihood of the data (LnP(X|*K*)) calculated over 20 STRUCTURE replicated runs is given for each value of the putative number of clusters (*K*). We used the admixture model with correlated allele frequencies and sampling location as prior information. The maximum value of LnP(X|*K*) is obtained for *K*=2. Note that the *ΔK* method (Evanno et al., 2005, *Molecular Ecology* 14:2611-2620) provides the same result (*K*=2).

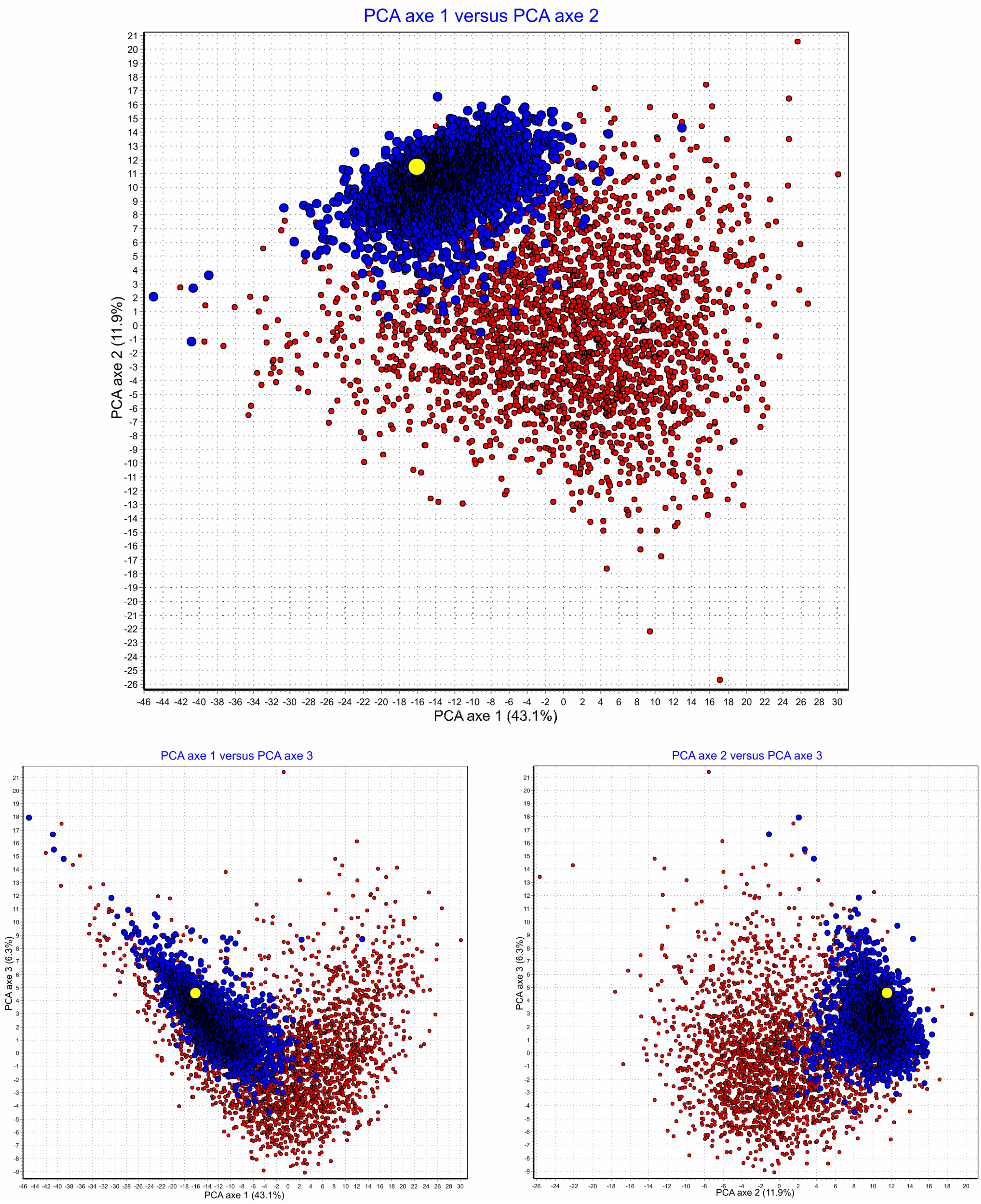

Figure S4: Graphical representation of the result of a principal component analysis (PCA) in the space of the summary statistics performed on the final selected worldwide invasion scenario.

In this PCA, the observations are the simulated data sets and the variables are the summary statistics. Each blue dot corresponds to a dataset simulated with parameters drawn from the posterior distributions (2500 dataset are randomly shown here). The yellow dot corresponds to the real *H. axyridis* dataset. Each red dot corresponds to a dataset simulated with parameters drawn from the prior distributions (2500 dataset are randomly shown here).

Population code name	Sampling location and historical information	Geographic coordinates	Sampling date (month-year)	Number of genotyped individuals
N-Russia1	Native area: Abakan, Khakassia, Russia	53.73°N 91.46°N	10-2007	31
N-Russia2	Native area: Novosibirsk, Novosibirsk Oblast, Russia	55.04°N 82.93°E	10-2007	30
N-Kazak	Native area: Almaty, Oblys d'Almaty, Kazakhstan	43.24°N 76.95°E	10-2007	26
N-China1 (§)	Native area: Beijing, China	40.24°N 116.23°E	05-2007	28
N-China2 (§)	Native area: Shilin City, Yunnan Province, China	24.90°N 103.35°E	08-2007	35
N-China3	Native area: Changchun City, Jilin Province, China	43.88°N 125.31°E	11-2006	29
N-Japan1 (§)	Native area: Fuchu, Japan	34.57°N 133.24°E	09-2005	36
N-Japan2	Native area: Kyoto, Japan	35.01°N 135.77°E	08-2008	26
N-Korea	Native area: Daejeon, South Korea	36.37°N 127.35°E	11-1998	30
EB-INRA87 (§ *)	European Biocontrol: Rearing stock, INRA laboratory History: descendent from a population sampled in China in 1982	-	04-1987	18
EB-INRA06	European Biocontrol: Rearing stock, INRA laboratory History: descendant of EB-INRA87	-	11-2006	27
EB-Biotop	European Biocontrol: Rearing stock, Biotop biofactory History: strain obtained by Biotop from EB-INRA87 in 1995	-	11-2007	29
EB-Koppert (*)	European Biocontrol: Rearing stock, Koppert biofactory History: strain obtained by Koppert from EB-Biotop in 1997	-	07-2003	20
EB-Biobest (*)	European Biocontrol: Rearing stock, Ghent University laboratory (obtained from Biobest biofactory in 2003) History: strain obtained by Biobest from EB-Biotop in 1997	-	04-2007	27
UB-US (*)	North American Biocontrol: Insect collection USDA (http://www.ars-grin.gov/cgi-bin/nigrp/robo/f941s.pl?50902)	-	10-1980	25
I-ENA (§)	Invasive area: Joyce, Louisiana, USA	31.94°N 92.60°W	11-2007	34
I-WNA (§)	Invasive area: Sunnyside, Washington, USA	46.32°N 120.01°W	09-2007	42
I-EU (§)	Invasive area: Gent, Belgium	51.05°N 3.71°E	10-2007	32
I-SA (§)	Invasive area: Curitiba, Brazil	25.45°S 49.24°W	02-2008	30
I-AF (§)	Invasive area: Somerset West, South Africa	34.03°S 18.83°E	05-2008	31

Table S1: Native, biocontrol and invasive population samples of *Harmonia axyridis* used in this study.

In the "Population code name" column, "§" indicates that the corresponding sample was previously used by Lombaert *et al.* (2010), and "*" indicates that the corresponding sample was stored dry and was thus difficult to genotype at some loci (see main text for details).

pop	N-China1	N-China2	N-China3	N-Japan1	N-Japan2	N-Korea	N-Russia1	N-Russia2	N-Kazak	EB-INRA87	EB-INRA06	EB-Koppert	EB-Biobest	EB-Biotop	UB-US	I-ENA	I-WNA	I-EU	I-SA
N-China2	0.010																		
N-China3	0.009	0.002																	
N-Japan1	0.007	0.000	0.005																
N-Japan2	0.007	0.006	**0.004**	**0.002**															
N-Korea	0.018	0.008	0.012	0.005	0.008														
N-Russia1	0.033	0.021	0.021	0.020	0.019	0.015													
N-Russia2	0.035	0.027	0.027	0.026	0.023	0.022	**-0.006**												
N-Kazak	0.024	0.015	0.013	0.015	0.012	0.014	**0.004**	**0.003**											
EB-INRA87	0.179	0.149	0.165	0.164	0.161	0.157	0.178	0.174	0.182										
EB-INRA06	0.261	0.216	0.240	0.225	0.231	0.230	0.262	0.277	0.271	0.231									
EB-Koppert	0.211	0.168	0.182	0.176	0.182	0.181	0.202	0.213	0.208	0.140	0.174								
EB-Biobest	0.319	0.305	0.309	0.308	0.318	0.319	0.313	0.312	0.325	0.243	0.381	0.312							
EB-Biotop	0.233	0.190	0.196	0.203	0.214	0.205	0.226	0.235	0.224	0.165	0.218	0.062	0.332						
UB-US	0.029	0.031	0.017	0.031	0.033	0.050	0.063	0.067	0.053	0.179	0.301	0.205	0.294	0.204					
I-ENA	0.029	0.017	0.012	0.016	0.022	0.034	0.036	0.041	0.028	0.188	0.255	0.207	0.335	0.224	0.048				
I-WNA	0.026	0.008	0.009	0.011	0.016	0.019	0.030	0.038	0.019	0.184	0.239	0.197	0.325	0.206	0.047	0.023			
I-EU	0.061	0.043	0.046	0.048	0.049	0.055	0.069	0.080	0.069	0.111	0.158	0.070	0.231	0.120	0.068	0.059	0.064		
I-SA	0.114	0.096	0.097	0.097	0.101	0.103	0.122	0.143	0.116	0.279	0.312	0.248	0.418	0.281	0.170	0.064	0.107	0.109	
I-AF	0.039	0.032	0.030	0.034	0.042	0.030	0.031	0.041	0.034	0.188	0.286	0.209	0.337	0.221	0.065	0.023	0.037	0.066	0.089

Table S2: Pairwise estimates of F_{ST} between all *Harmonia axyridis* population sample pairs.
F_{ST} in bold typeface indicates non significant pair-wise differentiation, as assessed in Fisher's exact test with correction for multiple comparisons.

		Competing scenario set (reference table)			
		Sampled origin (*SO*)		Unsampled origin (*UO*)	
Pods' Parameter distribution set	Scenario considered	Type I error	Type II error	Type I error	Type II error
Broad with *ar* = 0.5 and t_{anc} = 3000	SA	0.030	0.000	0.000	0.010
	SB	0.020	0.005	0.010	0.005
	SAB	0.010	0.025	0.030	0.005
	S mean	**0.020**	**0.010**	**0.013**	**0.007**
	UA	0.400	0.010	0.120	0.030
	UB	0.400	0.015	0.100	0.050
	UAB	0.050	0.400	0.130	0.095
	U mean	**0.283**	**0.142**	**0.117**	**0.058**

Table S3: Confidence in scenario selection based on ABC analyses on pseudo-observed data sets. Results obtained with (i) intermediate admixture rate and (ii) high splitting time between the two native population clusters.

The compared scenarios are detailed in Figure 2. Parameter distributions are given in Table 1, except for *ar* and t_{anc} which are here fixed to 0.5 and 3000 respectively. Type I error: proportion of cases in which the scenario considered is excluded but is actually the true one. Type II error: proportion of cases in which the scenario considered is selected but is not the true one.

		Competing scenario set used	
Biocontrol sample	Scenarios	Sampled origin (*SO*)	Unsampled origin (*UO*)
EB-INRA87	Western cluster	0.0191 [0.0117,0.0266]	0.2441 [0.1940,0.2943]
	Eastern cluster	***0.4924 [0.4347,0.5501]***	**0.4946 [0.4426,0.5465]**
	Admixture West/East	***0.4884 [0.4317,0.5452]***	0.2613 [0.2228,0.2998]
EB-INRA06	Western cluster	0.0081 [0.0040,0.0121]	0.0861 [0.0587,0.1135]
	Eastern cluster	**0.7527 [0.6974,0.8079]**	**0.6950 [0.6397,0.7503]**
	Admixture West/East	0.2392 [0.1852,0.2933]	0.2189 [0.1731,0.2647]
EB-Biotop	Western cluster	0.0022 [0.0010,0.0035]	0.0435 [0.0275,0.0596]
	Eastern cluster	**0.8852 [0.8574,0.9129]**	**0.7878 [0.7471,0.8284]**
	Admixture West/East	0.1126 [0.0852,0.1400]	0.1687 [0.1333,0.2041]
EB-Koppert	Western cluster	0.0012 [0.0006,0.0018]	0.0751 [0.0547,0.0954]
	Eastern cluster	**0.7888 [0.7521,0.8255]**	**0.6987 [0.6580,0.7394]**
	Admixture West/East	0.2100 [0.1734,0.2466]	0.2263 [0.1914,0.2611]
EB-Biobest	Western cluster	0.0328 [0.0226,0.0431]	0.1947 [0.1589,0.2305]
	Eastern cluster	**0.6047 [0.5620,0.6474]**	**0.5024 [0.4603,0.5444]**
	Admixture West/East	0.3625 [0.3215,0.4035]	0.3029 [0.2670,0.3389]
UB-US	Western cluster	0.0004 [0.0002,0.0006]	0.1047 [0.0816,0.1279]
	Eastern cluster	**0.7557 [0.7177,0.7938]**	**0.7470 [0.7163,0.7776]**
	Admixture West/East	0.2439 [0.2058,0.2819]	0.1483 [0.1287,0.1679]

Table S4: ABC posterior probabilities of the three competing scenarios modeling the genetic origin of each biocontrol sample within the native area of *Harmonia axyridis* (western native cluster, eastern native cluster or admixture of the western and eastern native clusters).

95% confidence intervals (CI) are shown in square brackets. Results are given for a sampled origin scenario design (*SO*) or an unsampled origin scenario design (*UO*); see main text for details. Posterior probabilities, with CI, of the selected scenarios are shown in bold typeface (and in bold/italic typeface when the CIs of several scenarios overlap).

	Scenarios	Posterior probabilities
S1	Native ➔ (EB-INRA87); EB-INRA87 ➔ (EB-INRA06, EB-Biotop, EB-Koppert, EB-Biobest)	0.999 [0.995,1.000]
S2	Native ➔ (EB-INRA87, EB-Koppert); EB-INRA87 ➔ (EB-INRA06, EB-Biotop, EB-Biobest)	0.000 [0.000,0.000]
S3	Native ➔ (EB-INRA87, EB-Biobest); EB-INRA87➔ (EB-INRA06, EB-Biotop, EB-Koppert)	0.001 [0.000,0.005]
S4	Native ➔ (EB-INRA87, EB-INRA06); EB-INRA87➔ (EB-Biotop, EB-Koppert, EB-Biobest)	0.000 [0.000,0.000]
S5	Native ➔ (EB-INRA87, EB-Biotop); EB-INRA87 ➔ (EB-INRA06, EB-Koppert, EB-Biobest)	0.000 [0.000,0.000]
S6	Native ➔ (EB-INRA87, EB-Koppert, EB-Biobest); EB-INRA87 ➔ (EB-INRA06, EB-Biotop)	0.000 [0.000,0.000]
S7	Native ➔ (EB-INRA87, EB-INRA06, EB-Koppert); EB-INRA87 ➔ (EB-Biotop, EB-Biobest)	0.000 [0.000,0.000]
S8	Native ➔ (EB-INRA87, EB-Biotop, EB-Koppert); EB-INRA87 ➔ (EB-INRA06, EB-Biobest)	0.000 [0.000,0.000]
S9	Native ➔ (EB-INRA87, EB-INRA06, EB-Biobest); EB-INRA87 ➔ (EB-Biotop, EB-Koppert)	0.000 [0.000,0.000]
S10	Native ➔ (EB-INRA87, EB-Biotop, EB-Biobest); EB-INRA87 ➔ (EB-INRA06, EB-Koppert)	0.000 [0.000,0.000]
S11	Native ➔ (EB-INRA87, EB-INRA06, EB-Biotop); EB-INRA87 ➔ (EB-Koppert, EB-Biobest)	0.000 [0.000,0.000]
S12	Native ➔ (EB-INRA87, EB-INRA06, EB-Koppert, EB-Biobest); EB-INRA87 ➔ (EB-Biotop)	0.000 [0.000,0.000]
S13	Native ➔ (EB-INRA87, EB-Biotop, EB-Koppert, EB-Biobest); EB-INRA87 ➔ (EB-INRA06)	0.000 [0.000,0.000]
S14	Native ➔ (EB-INRA87, EB-INRA06, EB-Biotop, EB-Koppert); EB-INRA87 ➔ (EB-Biobest)	0.000 [0.000,0.000]
S15	Native ➔ (EB-INRA87, EB-INRA06, EB-Biotop, EB-Biobest); EB-INRA87 ➔ (EB-Koppert)	0.000 [0.000,0.000]
S16	Native ➔ (EB-INRA87, unsampled pop); EB-INRA87 ➔ (EB-INRA06, EB-Biotop); unsampled pop ➔ (EB-Koppert, EB-Biobest)	0.000 [0.000,0.000]
S17	Native ➔ (EB-INRA87, unsampled pop); EB-INRA87 ➔ (EB-Koppert, EB-Biobest); unsampled pop ➔ (EB-INRA06, EB-Biotop)	0.000 [0.000,0.000]
S18	Native ➔ (EB-INRA87, EB-INRA06, EB-Biotop, EB-Koppert, EB-Biobest)	0.000 [0.000,0.000]

Table S5: ABC analyses to assess the relationship between the five European Biocontrol populations.
For formal assessment of the relationship between the European biocontrol populations, we used ABC to compare 18 competing scenarios representing of a gradient from a scenario of full dependence (i.e. all samples derived from the 1982 INRA population: S1) to a scenario of full independence (i.e. all samples independently collected within the native area: S18). For instance, in scenario S3, the biocontrol populations EB-INRA87 and EB-Koppert were independently collected from the native area and all three biocontrol populations EB-INRA06, EB-Biotop and EB-Koppert originate from the same biocontrol population, EB-INRA87. Building on the results we obtained considering each biocontrol

population separately (Table S4), we used the eastern native population cluster as the native source population for all the scenarios compared (we used N-China2 as the native sample) and the *UO* scenario set design. Parameter priors were those used in the “broad parameter distribution set” used in simulation analyses (Table 1), assuming 2.5 generations per year and with biocontrol populations assumed to maintain a low effective size that has remained constant over time since their collection (i.e. log uniform distribution [10;1000]). 5x10^5 microsatellite data sets per scenario were simulated. The steps of the ABC were as described in the main text, section “*ABC analyses on controlled simulated data sets*”. Posterior probability of each scenario is given with 95% confidence interval between brackets.

We found, with a very high posterior probability (P = 0.999; 95% CI = 0.995 – 1.000) that all European biocontrol strains were derived from the same population (i.e. S1 = full dependence scenario). We also performed a model checking analysis under the selected scenario (S1). To do so, 10^6 data sets were simulated. A “posterior sample” of 10^4 values of the posterior distributions was obtained. We then simulated 10^4 data sets with parameter values drawn, with replacement, from this “posterior sample”. Our set of test statistics were the one described in the main text. We found that only 8 statistics out of 144 were in the tail of the distributions of the statistics simulated from the posterior predictive distributions. Altogether, these results confirmed that the main biofactories in Europe had been rearing *H. axyridis* samples originating from the same population collected by INRA in the eastern part of the native area in 1982.

Parameter	Distribution	Mean	Median	Mode	Quantile 2.5%	Quantile 97.5%
NS_i and NS_j	Uniform [100 – 20,000]	10,056	10,040	NA	640	19,490
NS_k	Loguniform [10 – 1,000]	506	508	NA	35	975
NF_i	Loguniform [2 – 1,000]	162	45	2	2	862
BD_i	Uniform [0 – 5]	2.5	2.5	NA	0	5
ar	Uniform [0.1 – 0.9]	0.5	0.5	NA	0.12	0.88
t_i	Uniform [x_i – x_i+5]	DV	DV	NA	DV	DV
tbc_i	Loguniform [t_i – 93]	DV	DV	DV	DV	DV
tu_j	Loguniform [tbc_i – 3000]	DV	DV	DV	DV	DV
t_{anc}	Uniform [100 – 3000]	1,858	1,940	NA	380	2,960
mean μ	Uniform [10^{-5} – 10^{-3}]	$5.0\text{x}10^{-4}$	$5.0\text{x}10^{-4}$	NA	$3.5\text{x}10^{-5}$	$9.8\text{x}10^{-4}$
mean *P*	Uniform [0.1 – 0.3]	0.2	0.2	NA	0.10	0.29
mean μSNI	Uniform [10^{-8} – 10^{-4}]	$5.0\text{x}10^{-5}$	$5.0\text{x}10^{-5}$	NA	$2.5\text{x}10^{-6}$	$9.7\text{x}10^{-5}$

Table S6: Prior distributions of demographic, historic and mutation parameters used in ABC analyses attempting to retrace the worldwide routes of invasion of *Harmonia axyridis*.

Notes: Populations *i* are invasive populations, clusters *j* are native clusters (either western or eastern cluster) and populations *k* correspond to biocontrol strains (i.e. laboratory reared populations). Times were translated into numbers of generations running back in time and assuming 2.5 generations per year. *NS* = stable effective population size (number of diploid individuals); *NF* = effective number of founders during an introduction step lasting *BD* generation(s); *ar* = admixture rate (only for scenarios with admixture); t_i = introduction date of invasive populations *i* with limits x_i fixed from dates of first observation, assuming 2.5 generations per year; *tbc* = creation date of unsampled biocontrol strain for eastern and western North American populations bounded by the dates of the first observation of the invasive population (corresponding to a direct introduction into the wild) and the number of generations from 1970, the start date of a period of intense *H. axyridis* biocontrol activity in the USA; tu_j = in native cluster j, date of merging of the source unsampled native population with the sampled native population (this parameter is included only in the model in which the scenario contains one or both native populations as possible source(s)); t_{anc} = date of the merging of the two native populations into an ancestral unsampled population (with condition $tu_j \leq t_{anc}$). For microsatellite marker parameters, parameters were as in Table 1. All prior quantities presented were calculated from 100,000 values. NA = not applicable; DV = may take different values.