Colonization history of the Western Corn Rootworm (*Diabrotica virgifera virgifera*) in North America: insights from random forest ABC using microsatellite data

Eric Lombaert[1], Marc Ciosi[2], Nicholas J. Miller[3], Thomas W. Sappington[4], Aurélie Blin[1] and Thomas Guillemaud[1]

[1] Université Côte d'Azur, INRA, CNRS, ISA, France
[2] Institute of Molecular Cell and Systems Biology, University of Glasgow, Glasgow, UK
[3] Department of Biology, Illinois Institute of Technology, 3101 S Dearborn St, Chicago, Illinois 60616, USA
[4] USDA-Agricultural Research Service, Corn Insects & Crop Genetics Research Unit, Genetics Laboratory, Iowa State University, Ames, Iowa, USA

Keywords: biological invasion, invasion routes, approximate Bayesian computation, maize.

Corresponding author:
Eric Lombaert – Institut Sophia Agrobiotech – 400, route des chappes – BP 167 – 06903 Sophia-Antipolis Cedex - France
E-mail: lombaert@sophia.inra.fr
Tel: +33 4 92 38 65 06
Fax: +33 4 92 38 64 01

Running title: Biogeography of the Western Corn Rootworm in America

Abstract

First described from western Kansas, USA, the western corn rootworm, *Diabrotica virgifera virgifera*, is one of the worst pests of maize. The species is generally thought to be of Mexican origin and to have incidentally followed the expansion of maize cultivation into North America thousands of years ago. However, this hypothesis has never been investigated formally. In this study, the genetic variability of samples collected throughout North America was analysed at 13 microsatellite marker loci to explore precisely the population genetic structure and colonization history of *D. v. virgifera*. In particular, we used up-to-date Approximate Bayesian Computation methods based on Random Forest algorithms to test a Mexican versus a central-USA origin of the species, and to compare various possible timings of colonization. This analysis provided strong evidence that the origin of *D. v. virgifera* was Mexico, or even further south. Surprisingly, we also found that the expansion of the species north of its origin was recent – probably starting less than 1000 years ago – thus indicating it was not directly associated with the history of maize expansion out of Mexico, a far more ancient event.

Introduction

The western corn rootworm (WCR), *Diabrotica virgifera virgifera*, is a major economic pest of maize, *Zea mays*, in North America and, since the end of the twentieth century, in Europe (Gray et al., 2009; Vilà et al., 2009). Although the invasion history of WCR in Europe has been well investigated (Miller et al., 2005; Ciosi et al., 2008), its biogeography, colonisation history and potential association with maize domestication in America are poorly understood.

Because of the geographical distribution of most other diabroticites and the close association of WCR with maize, the species is commonly considered as originating from Mexico, or possibly Guatemala, where its original native host was probably *Tripsacum*, a close wild relative to maize (Smith, 1966; Branson & Krysan, 1981; Gray et al., 2009). The classically proposed scenario is that WCR fed on early domesticated maize, and that it has incidentally followed the dissemination of the plant all over North America so that the history of WCR tracks the history of maize (Branson & Krysan, 1981). Maize is a human-made variant of teosinte which was domesticated between 12,000 and 8,700 years before present (BP) in southern Mexico. The cultivation of maize slowly expanded northward to reach the present-day states of Arizona and New Mexico, USA, between 4,500 and 3,000 years BP. The selection of new variants that were better adapted to temperate climates helped to spread maize further into present-day USA and Canada around 2,000 years BP. A large increase in maize cultivation by European migrants in North America occurred in the nineteenth century, probably helped by a selection of new cultivars. Finally, the intensification of cultivation after 1950 widely boosted this trend (Doebley et al., 1988; Fritz, 1990; Boyd et al., 2008; Merrill et al., 2009; Tenaillon & Charcosset, 2011; Hufford et al., 2012).

However, different WCR origin scenarios are possible and have been proposed, such as a far more recent colonization history than that of maize, and/or a more northern North American origin of the species (Chiang, 1973; Gray et al., 2009). These scenarios are based on the dates of first observation of WCR in America and on the knowledge of its ecology. *D. virgifera* was first described by Le Conte from two individuals collected in 1867 from blossoms of *Cucurbita foetidissima* in western Kansas (Le Conte, 1868; Metcalf, 1983; Krysan & Smith, 1987), and the first economic damage on maize was noticed only in 1909 in Colorado (Gillette, 1912). The species is known to have been present in more southern States such as Arizona and New Mexico, as well as in Mexico, at least since the end of the nineteenth century (Horn, 1893), but more detailed information about their presence in these areas was not available before the 1950s (Chiang, 1973; Krysan & Smith, 1987). The colonization of the Eastern USA and Canada by WCR has been well monitored and is very recent compared to the cultivation of maize in those areas: beginning in the 1940s, WCR started to spread eastward at considerable speed across North America to reach the East coast in the mid-1980s (Krysan & Smith, 1987; Gray et al., 2009; Meinke et al., 2009). Furthermore, behavioural data do not fully support an exclusive shared history between WCR and maize, suggesting instead a switch, which could possibly be recent, from a very different host plant (than *Tripsacum*) to maize, either in Mexico or the central USA. Indeed,

larvae have no mechanism for distinguishing maize from a distance (Branson & Krysan, 1981), whereas WCR adults are strongly attracted to cucurbitacins, secondary metabolites of Cucurbitaceae (Metcalf & Lampman, 1989). Potential alternative hosts in North America include a number of native grass species (Clark & Hibbard, 2004; Oyediran et al., 2004), but their current importance in a maize-dominated agroecosystem is probably minimal (Moeser & Hibbard, 2005; Campbell & Meinke, 2006).

In this study, we characterized the current genetic structure of WCR in North America, from Mexico to the north-eastern USA, by Bayesian clustering methods and more classical population genetic statistics and methods. We then performed approximate Bayesian computation analyses to quantitatively compare colonization scenarios of WCR populations in North America.

Methods

Sampling, genotyping and genetic variation

917 WCR adults were collected from 21 sites (14 to 62 WCR per site) in North America between 1998 and 2006, covering a substantial part of the distribution of this species in America (Fig. 1; Table S1). Samples from twelve of these sites were used in previous studies (Table S1; Kim & Sappington, 2005; Kim et al., 2008; Coates et al., 2009). Genotyping at 13 microsatellite marker loci was carried out in three separate multiplex PCRs for all individuals as described by Bermond *et al.* (2012).

Genetic variation within and between the 21 site-samples were quantified by calculating the mean number of alleles per locus *NA*, the mean expected heterozygosity H_e (Nei, 1987) and pairwise F_{ST} estimates (Weir & Cockerham, 1984) using Genepop (version 4.2, Raymond & Rousset, 1995). To take into account the differences in sample size, for 20 site-samples we computed the mean allelic richness (*AR*) corrected for 18 individuals by the rarefaction method (Petit et al., 1998) with FSTAT (version 2.9.3.2, Goudet, 1995). Hardy-Weinberg and genotypic differentiation tests were performed using Fisher exact tests implemented in Genepop (version 4.2, Raymond & Rousset, 1995), and significance levels were corrected for multiple comparisons biases by the false discovery rate procedure (Benjamini & Hochberg, 1995). We constructed a neighbour-joining (NJ) tree (Saitou & Nei, 1987) using pairwise genetic distances as described by Cavalli-Sforza and Edwards (1967), using Populations software (version 1.2.30, Langella, 1999). The robustness of tree topology was evaluated by carrying out 1000 bootstrap replicates over loci.

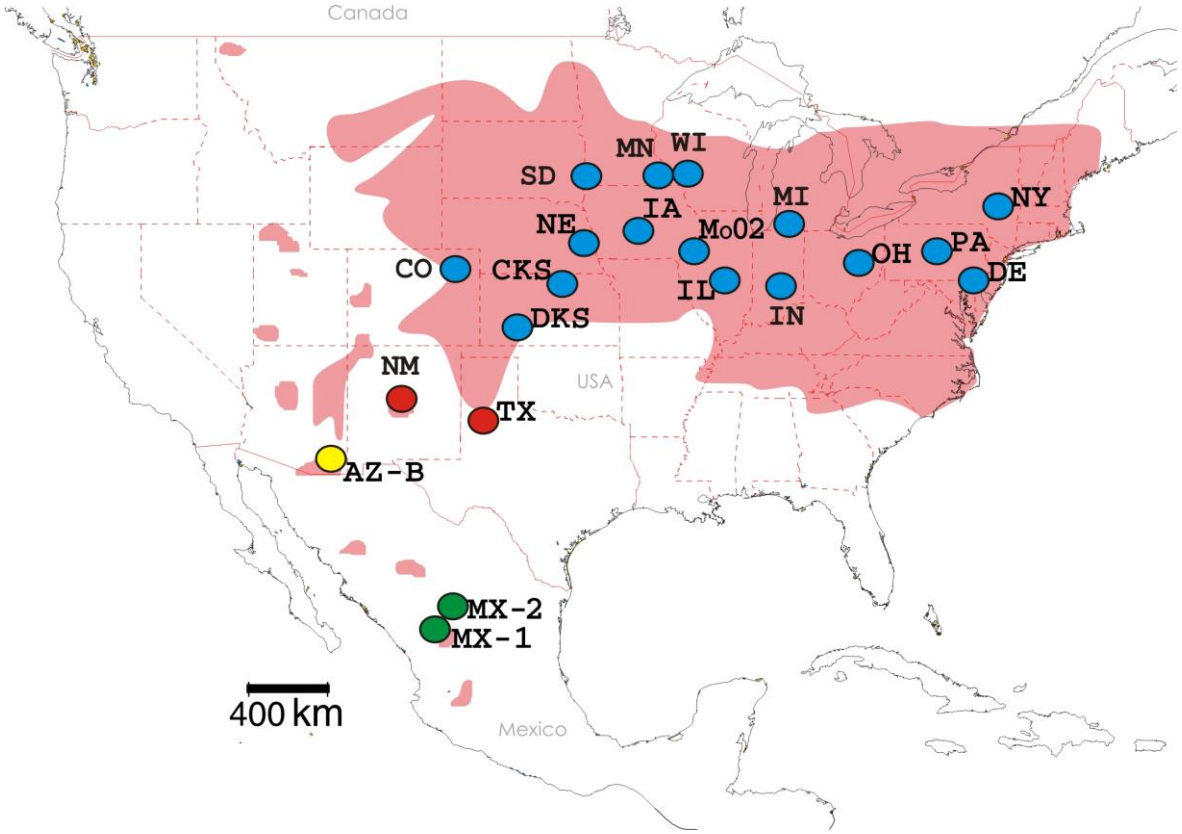

Figure 1: Geographic locations of genotyped site-samples of WCR and genetic units inferred from Bayesian clustering analyses.
Notes: Site-sample names are as in Table S1. The pink areas roughly correspond to the geographic distribution of WCR in North America. Site-samples of the same color belong to the same genetic unit, as assessed by hierarchical procedures applied to the Bayesian clustering methods implemented in STRUCTURE and BAPS (Figures S2 and S3): “Mexico” in green, “Arizona” in yellow, “New Mexico/Texas” in red and “Colorado/New York” in blue.

Population structure and definition of genetic units

The clustering approach implemented in STRUCTURE (v2.3.4, Pritchard et al., 2000) was used to infer the number of potential genetic units within the North American range of WCR. We chose the admixture model with correlated allele frequencies, and default values for all other parameters of the software. Each run consisted of a burn-in period of 2×10^5 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, with K set between 1 and [the number of site-samples considered + 1]. To group each site-sample within its most likely genetic unit, we used the hierarchical approach of Coulon *et al.* (2008) as follows. We first analysed the whole dataset, consisting of 21 site-samples (totalling 917 individuals). If the mean natural logarithm of the likelihood of the data $ln(P(X|K))$ was maximal for $K = 1$, then the inferred number of clusters was 1 and we stopped the procedure. Otherwise, we determined the best value of K by the ΔK method (Evanno et al., 2005). We then assigned each site-sample to the cluster for which the mean individual inferred ancestry was highest, provided this value was > 0.8; site-samples with maximum inferred ancestry < 0.8 were assigned to a specific “admixed” group. We repeated new STRUCTURE analyses

independently within each inferred cluster until *K* = 1 was the most likely, or until only one site-sample remained.

We also used the clustering approach implemented in BAPS software (v5.2, Corander et al., 2003) as a complement to the STRUCTURE analyses. BAPS analyses were carried out on groups of individuals (i.e. site-samples) rather than individuals, with simple model assumptions (i.e. no admixture and uncorrelated allele frequencies). We conducted a series of 20 replicate runs, with the upper limit for the number of clusters set as the actual number of sampled sites. BAPS infers the number of clusters (*K* is a parameter of the model), but we proceeded to a hierarchical approach as well, by performing independent analyses within each inferred cluster until the number of newly inferred clusters was one or until only one site-sample remained.

ABC-based inferences about colonization history

An approximate Bayesian computation analysis (ABC; Beaumont et al., 2002) was carried out to infer the colonization history of WCR in North America. The populations considered in the ABC analysis corresponded to the genetic units previously identified by the two Bayesian clustering methods (i.e. STRUCTURE and BAPS), and each genetic unit was represented in the analysis by a single site-sample (the "core dataset", see Results section). ABC is a model-based Bayesian method allowing posterior probabilities of historical scenarios to be computed, based on genetic and historical data. The history of maize cultivation along with the areas and dates of first observations of WCR in North America were used to define 6 competing colonization scenarios differing in the combination of three main characteristics. First, the geographical origin of the species: WCR either originated from Mexico and expanded northward ("Mexican origin"), or it originated near present-day Colorado and expanded southward and eastward ("central-USA origin"). Second, the demographic history of the scenario's first colonizing population: this population experienced either an "ancient bottleneck" (between 10,000 and 1,500 years BP) or a "recent bottleneck" (between 1,500 years BP and the date of first observation). This bottleneck could be the signal either of an introduction event from a native, unsampled, population or of a sudden decrease in population size during a selective sweep due to host plant shift. Third, the dates of the colonization events: either WCR accompanied the North American expansion of maize ("ancient expansion", between 10,000 years BP and 1,500 years BP), or its range expanded only recently ("recent expansion", between 1,500 years BP and the date of first observation). The competing scenarios thus differ in the direction of the colonization (south to north, or north to south) and by the relative recency of demographic and divergence events. In all scenarios, an expansion event corresponds to a simple divergence event from a source population possibly followed by a period at low effective size (bottleneck event) predating demographic stabilization at a higher effective size. Migration between all pairs of populations can occur unsymmetrically. All 6 scenarios are described in Table 1 and Figure S1.

In our ABC analysis, historical, demographic and mutational parameter values for simulations were drawn from prior distributions defined from historical data and from a previous study (Miller et al., 2005), as described in Table S2. We used a total of 49 summary statistics: for each population (i.e. site-sample in the case of the observed dataset), we computed the mean number of alleles per locus, the mean expected heterozygosity (Nei, 1987), the mean number of private alleles per locus and the mean ratio of the number of alleles to the range of allele sizes (Garza & Williamson, 2001). For each pair of populations, we computed the pairwise F_{ST} values (Weir & Cockerham, 1984) and the mean likelihoods of individuals from population *i* being assigned to population *j* (Rannala & Mountain, 1997). For each trio of populations we computed the maximum likelihood estimate of admixture proportion (Choisy et al., 2004). For all populations taken together, we computed the mean number of alleles per locus, the mean expected heterozygosity and the mean number of shared alleles per locus. These statistics were complemented with the five axes obtained from a linear discriminant analysis on summary statistics (Estoup et al., 2012).

To compare the scenarios, we used a random forest process (Breiman, 2001) as described by Pudlo *et al.* (2016). Random forest is a machine-learning algorithm which circumvents curse of dimensionality problems and some problems linked to the choice of summary statistics (e.g. correlations between statistics). This non-parametric classification algorithm uses hundreds of bootstrapped decision trees (creating the so-called forest) to perform classification using a set of predictor variables, here the summary statistics. Some simulations are not used in tree building at each bootstrap (i.e. the out-of-bag simulations) and can thus be used to compute the "prior error rate", which provides a direct method for cross-validation. Random forest (i) has large discriminative power, (ii) is robust to the choice and number of summary statistics and (iii) is able to learn from a relatively small reference table hence allowing a drastic reduction of computational effort. We simulated 50,000 microsatellite datasets for each competing scenario, and checked whether the scenarios and priors were off target or not by comparing distributions of simulated summary statistics with the value of the observed dataset. We then grew a classification forest of 1,000 trees based on all simulated datasets. The random forest computation applied to the observed dataset provides a classification vote which represents the number of times a model is selected among the 1,000 decision trees. The scenario with the highest classification vote was selected as the most likely scenario. We then estimated its posterior probability by way of a second random forest procedure of 1,000 trees as described by Pudlo *et al.* (2016). To evaluate the global performance of our ABC scenario choice, we (i) computed the *prior error rate* based on the available *out-of-bag* simulations, and (ii) conducted the scenario selection analysis a second time with another set of site-samples (the "alternative dataset") representative of the same genetic units as the core dataset, as suggested by Lombaert et al. (2014). Finally, we inferred posterior distribution values of all parameters, and some relevant composite parameters, of the selected scenario under a regression by random forest methodology (Marin et al., 2016), with classification forests of 1,000 trees.

Because the various populations under scrutiny are not separated by insurmountable geographical barriers, we allowed continuous migration between populations. We thus used ABCsampler (Wegmann et al., 2010) coupled with fastsimcoal2 (v2.5, Excoffier et al., 2013) for simulating datasets and generating reference tables. We used Arlequin 3.5 (using the arlsumstat console version, Excoffier & Lischer, 2010), in-house codes (perl and C++) and an R script used by Benazzo *et al.* (2015) to compute summary statistics. Scenario comparisons and parameter estimations were performed under R (R Development Core Team, 2015) with the "*abcrf*" package (v1.3, Pudlo et al., 2016).

Results

Genetic variation in WCR

The complete dataset, including a total of 917 individuals from 21 site-samples, displayed substantial polymorphism, with a mean of 12.69 alleles per locus, over all samples. Allelic richness corrected for 18 individuals ranged from 5.54 alleles per locus in a sample from Minnesota (MN) to 6.48 in a Mexican sample (MX-2). Overall, the southernmost site-samples displayed the highest diversities, especially in Mexico, and to a lesser extent in Arizona, New Mexico and Texas. See Table S1 for a concise presentation of diversity measurements for each site-sample.

Genotypic differentiation was statistically significant in 137 of 210 pairwise comparisons between site-samples (Table S3). Global levels of differentiation between site-samples were moderate, with a mean F_{ST} of 0.035. As previously described in other studies using lower numbers of samples and genetic markers (Kim & Sappington, 2005; Ciosi et al., 2008; Kim et al., 2008; Coates et al., 2009), a large part of the northern USA, i.e. all site-samples above the states of New Mexico and Texas, displayed high genetic similarity with a mean pairwise F_{ST} of 0.005. In contrast, F_{ST} values increased steeply with latitude, with the highest value (0.16) between site-samples MX-2 in Mexico and Mo-02 in Illinois (Table S3).

In the unrooted NJ tree, the position of the site-samples was mostly consistent with a latitudinal pattern (Fig. 2). Despite long branches, both Mexican samples grouped together, and were closest to Arizona, followed by New Mexico and Texas. The remaining 16 site-samples grouped together in a tight cluster with short branches.

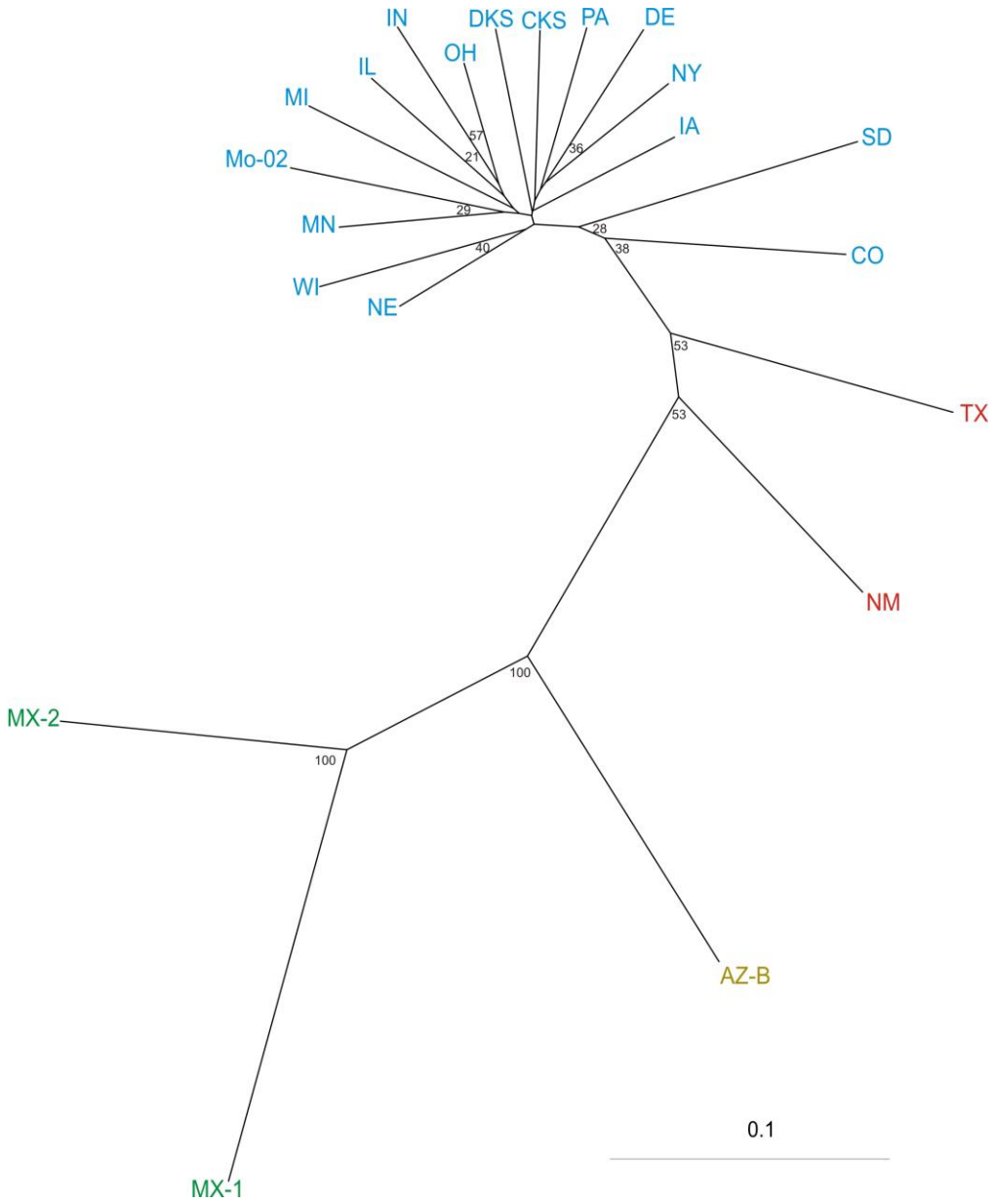

Figure 2: Neighbour-joining tree for WCR site-samples based on the chord distance of Cavalli-Sforza & Edwards (1967). Site-sample names are as in Figure 1 and Table S1. Site-samples of the same color belong to the same genetic unit as inferred from STRUCTURE and BAPS (Figures S2 and S3). Bootstrap values calculated over 1000 replications are given as percentages (only values >20% are shown).

Population structure of WCR in North America

A hierarchical approach applied to both STRUCTURE (Pritchard et al., 2000) and BAPS (Corander et al., 2003) Bayesian clustering methods provided the same qualitative results. In the first round, site-samples were partitioned into three groups: the first contained MX-1, MX-2 and AZ-B site-samples, the second contained the NM and TX site-samples, and the third contained all 16 remaining site-samples. Second rounds within each of these three groups only separated the two Mexican site samples (MX-1 and MX-2) from Arizona's single site-sample (AZ-B). Details of BAPS and STRUCTURE results can be found in Figures S2 and

S3. To summarize, our 21 site-samples could be partitioned into four main genetic units clearly linked to geographical patterns (Fig. 1): (i) the "Mexico" genetic unit (46 individuals from 2 site-samples: MX-1 and MX-2), (ii) the "Arizona" genetic unit (40 individuals from 1 site-sample: AZ-B), (iii) the "New Mexico/Texas" genetic unit (82 individuals from 2 site-samples: NM and TX) and (iv) the "Colorado/New York" genetic unit (749 individuals from 16 site-samples: CO, DKS, CKS, NE, SD, IA, MN, WI, Mo-02, IL, IN, MI, OH, PA, DE and NY).

Colonization history of WCR in North America inferred from ABC analyses

For the core dataset used in the ABC analyses, the choice of site-samples was based on the largest sample sizes for the "Mexico" and "New Mexico/Texas" genetic units: MX-2 and TX respectively. For the "Colorado/New York" genetic unit, we chose the site-sample CO from Colorado, because of its geographical proximity to the historical first observation of the species, and because of the well-described colonization history of this genetic unit eastward from this area (Gray et al., 2009). For the alternative dataset, the "Mexico" and the "New Mexico/Texas" genetic units were represented by the MX-1 and NM site-samples respectively, and the "Colorado/New York" genetic unit was represented by the OH site sample which displayed the lowest mean intra-genetic unit pairwise F_{ST} (Table S3). In both datasets, the "Arizona" genetic unit was represented by the single AZ-B site-sample. Regarding the clear geographical partition of the four genetic units (Fig. 1), and the patterns observed in the NJ tree (Fig. 2), the "Mexican origin" scenarios represent a simple South to North expansion in this specific order: (i) "Mexico", (ii) "Arizona", (iii) "New Mexico/Texas" and (iv) "Colorado/New York". The "central-USA origin" scenarios entail an expansion in the opposite direction, from North to South (Fig. S1). Raw dates of first observation were used as lower bounds of time prior distributions (Table S2): 1893 for "Mexico", "Arizona" and "New Mexico/Texas" (i.e. 113 generations backward in time, Horn, 1893), and 1867 for "Colorado/New York" (i.e. 139 generations back in time, Le Conte, 1868). Depending on the topology of the scenario, these dates were narrowed by conditions.

Comparisons of distribution of simulated summary statistics with values of the observed core dataset showed that the combination of scenarios and prior that we chose was realistic: among the six simulated scenarios, we had from zero (scenarios 1 and 5) to only two (scenarios 2, 4 and 6) observed statistics out of 49 that significantly (at a 5% threshold) lay in the tails of the probability distribution of statistics calculated from prior simulations (Table S4).

The results of the ABC analyses are shown in Table 1, and the selected scenario is graphically summarized in Figure 3. The results indicate, with a high probability of almost 0.7 for scenario 1, that (i) Mexico is the most likely first identifiable source of the colonization, (ii) a bottleneck occurred recently in this population and (iii) the colonization of North America by WCR is recent. The prior error rate was high (49.2%), but the result was qualitatively and quantitatively confirmed by the analysis of the alternative dataset which

selected the same scenario with a very similar posterior probability (Table 1). This high prior error rate was caused by some scenarios being differentiated only by the prior distribution of divergence times. Indeed, the three "Mexican origin" scenarios (i.e. scenarios 1, 3 and 5; Fig. S1) brought together a total of 962 votes among the 1000 generated decision trees, with scenario 5 (i.e. ancient ancestral bottleneck and recent colonization) garnering the second highest number of votes. When comparing in a new analysis only the 3 scenarios with a Mexican origin differing by the times of colonization (scenarios 1, 3 and 5), scenario 1 with all historical events being recent obtained 719 votes among 1000.

				Random Forest votes		Posterior probability	
Scenario	Origin of WCR	Demographic history of oldest population	Time of colonization	Core dataset	Alternative dataset	Core dataset	Alternative dataset
S1	**Mexico**	**Recent bottleneck**	**Recent expansion**	**644**	**684**	**0.6901**	**0.6766**
S2	USA	Recent bottleneck	Recent expansion	29	9	-	-
S3	Mexico	Ancient bottleneck	Ancient expansion	19	12	-	-
S4	USA	Ancient bottleneck	Ancient expansion	3	5	-	-
S5	Mexico	Ancient bottleneck	Recent expansion	299	287	-	-
S6	USA	Ancient bottleneck	Recent expansion	6	3	-	-

Table 1: Description of the competing scenarios and results of the ABC analyses to infer the colonization history of WCR. Results are provided for both core and alternative datasets. The line in bold characters corresponds to the selected (most likely) scenario.

Point estimates of key parameters from scenario 1 are presented in Figure 3 (complete results in Table S5). The "Mexico" genetic unit suffered a strong initial bottleneck probably around 1,000 years ago. The geographic expansion that followed northward was accompanied by successive bottlenecks of lesser severity than the ancestral one. Effective population size was lowest for the "Colorado/New York" genetic unit (median value of N_4 = 3,395 individuals) which is the more recent population. In contrast, the "New Mexico/Texas" genetic unit displayed the largest population size (median value of N_3 = 15,384 individuals). This geographically central population received the largest number of migrants from each of the three other genetic units (from 3.3 to 3.8 effective migrants per generation). Effective migration between genetic units was, however, globally moderate over North America (mean of all median effective number of migrants = 2 individuals per generation). Note that most parameter posterior distributions displayed large ranges (Table S5), so these results should be interpreted with caution.

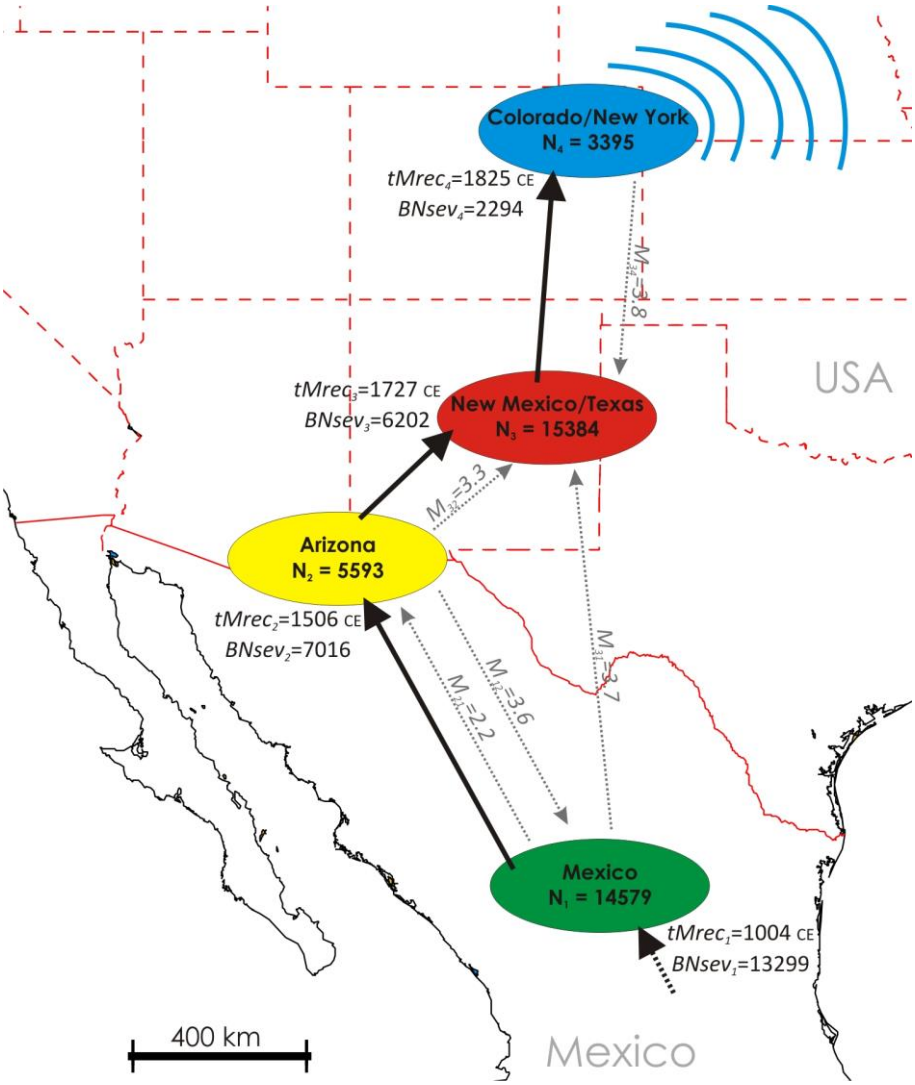

Figure 3: Graphical representation of the most likely scenario of WCR colonization of North America, and main parameter estimations.
Notes: The four genetic units are those inferred from Bayesian clustering analyses. All parameter estimations were performed with samples MX-2, AZ-B, NM and CO representing the "Mexico", "Arizona", "New Mexico/Texas" and "Colorado/New York" genetic units, respectively. All displayed parameter values are the medians of posterior distributions (Table S5). $BNsev_i$ = bottleneck severity of population *i* computed as $[BD_i \times N_{parental\ population\ of\ population\ i} / NF_i]$. M_{ij} is the effective number of migrants per generation from population *i* to population *j* backward in time, computed as $m_{ij} \times N_i$; only values above 2 individuals per generation are presented. All arrows are presented forward in time for ease of reading. Dates are presented in years of the Common Era (i.e. CE). Blue lines near the "Colorado/New York" genetic unit represent the well described eastward expansion after the 1940s (Gray et al., 2009).

Discussion

The main results of our study are that the origin of WCR is in the south of its North American range, and that it has expanded northward. ABC results were indeed confirmed by those of more classical population genetics methods, such as the observation of a decrease in genetic variation from South to North, as expected from successive founder events during a range expansion (Le Corre & Kremer, 1998; Hallatschek & Nelson, 2008). This quantitative

approach confirms what was previously proposed based on historical or phylogenetic data and rejects the hypothesis of a northern origin of WCR (Chiang, 1973; Branson & Krysan, 1981; Krysan & Smith, 1987; Gray et al., 2009). However, our data do not allow us to determine the precise origin of the species. Our Mexican samples were collected in the state of Durango, while the WCR actually may have originated from further south in the country, or even in Guatemala. Indeed, the estimated strong ancestral bottleneck could be the signature of a first colonization step from an unsampled ancestral population.

Another important and unexpected conclusion of our study is that the history of WCR colonisation of North America is not associated with the history of maize expansion out of Mexico. First, the initial severe bottleneck detected in the Mexican sample may be the signature of a very recent change of host from an unknown plant to maize, thousands of years after maize was domesticated. Note however that it is not yet possible to differentiate this hypothesis from a signal of expansion. Our data firmly indicate that WCR only recently started to expand northward from Mexico, less than 1,000 years ago, when maize had already been cultivated for more than one millennium as far north as Canada (Tenaillon & Charcosset, 2011). Our analysis suggests that the most recent WCR population around Colorado may have originated from colonization northward from New Mexico/Texas in the first half of the nineteenth century (Fig. 3). Finally, the absence of genetic structure that we observed from Colorado to New York is entirely consistent with the very recent subsequent colonization history by the species throughout this large area of great economic importance. This corroborates historical records (Chiang, 1973; Metcalf, 1983; Gray et al., 2009) and previous population genetics studies (Kim & Sappington, 2005; Ciosi et al., 2008; Kim et al., 2008; Coates et al., 2009). It also explains the low estimated effective population size of the "Colorado/New York" genetic unit despite large population densities in the field, which is consistent with a still unmet mutation-drift equilibrium.

In this paper, we have provided quantitative evidence for the first time of the southern origin of WCR in North America. Moreover, our results strongly suggest that the colonization of WCR in North America is very recent. Thus it appears that the species was not gradually co-domesticated with maize, but rather behaved as an invasive species. From its tropical origin, the species has quickly adapted to continental climates and has become one of the worst pests of maize. Considering the estimated chronology of the North American invasion, and the very likely underlying association with key modifications of maize cultural practices, WCR can clearly be considered a product of modern agriculture, i.e. a recent man-made pest (Metcalf, 1986).

Acknowledgments

We thank our colleagues Rosanna Giordano, Stephan Toepfer, Uwe Stoltz, Kyung Seok Kim, Sue Ratcliff, Greg Cronholm, Lee French, Lance Meinke, Brendon Reardon, Eli Levine, Bruce Eisley, Dennis Calvin, Joanne Whalen and Ken Wise for *Diabrotica virgifera virgifera* beetles and DNA samples. We also thank Emeline Deleury, Arnaud Estoup and Andrea Benazzo for scripts to compute summary statistics. We also thank Alexandra Auguste, Paulette Flacchi and Marie-José Odonne for technical and administrative assistance. This work was funded by grants from ANR projects Bioinv4I and Emile, and from the French Agropolis Fondation (Labex Agro-Montpellier, BIOFIS).

References

Beaumont M.A., Zhang W.Y., & Balding D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Benazzo A., Ghirotto S., Vilaça S.T., & Hoban S. (2015) Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source. *Heredity*, 262–272.

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.

Bermond G., Ciosi M., Lombaert E., Blin A., Boriani M., Furlan L., Toepfer S., & Guillemaud T. (2012) Secondary contact and admixture between independently invading populations of the Western Corn Rootworm, Diabrotica virgifera virgifera in Europe. *PLoS One*, **7**, 12.

Boyd M., Varney T., Surette C., & Surette J. (2008) Reassessing the northern limit of maize consumption in North America: stable isotope, plant microfossil, and trace element content of carbonized food residue. *Journal of Archaeological Science*, **35**, 2545–2556.

Branson T.F. & Krysan J.L. (1981) Feeding and oviposition behavior and life cycle strategies of *Diabrotica*: an evolutionary view with implications for pest management. *Environmental Entomology*, **10**, 826–831.

Breiman L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Campbell L.A. & Meinke L.J. (2006) Seasonality and adult habitat use by four *Diabrotica* species at prairie-corn interfaces. *Environmental Entomology*, **35**, 922–936.

Cavalli-Sforza L.L. & Edwards A.W.F. (1967) Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics*, **19**, 233–257.

Chiang H.C. (1973) Bionomics of the northern and western corn rootworms. *Annual Review of Entomology*, **18**, 47–72.

Choisy M., Franck P., & Cornuet J.M. (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology*, **13**, 955–968.

Ciosi M., Miller N.J., Kim K.S., Giordano R., Estoup A., & Guillemaud T. (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.

Clark T.L. & Hibbard B.E. (2004) Comparison of nonmaize hosts to support western corn rootworm (Coleoptera: Chrysomelidae) larval biology. *Environmental Entomology*, **33**, 681–689.

Coates B.S., Sumerford D. V, Miller N.J., Kim K.S., Sappington T.W., Siegfried B.D., & Lewis L.C. (2009) Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *Journal of Heredity*, **100**, 556–564.

Le Conte J.L. (1868) New Coleoptera collected on the survey for the extension of the Union Pacific Railway, E. D. from Kansas to Fort Craig, New Mexico. *Transactions of the American Entomological Society*, **2**, 49–59.

Corander J., Waldmann P., & Sillanpaa M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Le Corre V. & Kremer A. (1998) Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model. *Journal of Evolutionary Biology*, **11**, 495–512.

Coulon A., Fitzpatrick J.W., Bowman R., Stith B.M., Makarewich C.A., Stenzler L.M., & Lovette I.J. (2008) Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma coerulescens*). *Molecular Ecology*, **17**, 1685–1701.

Doebley J., Wendel J.D., Smith J.S.C., Stuber C.W., & Goodman M.M. (1988) The origin of cornbelt maize: The isozyme evidence. *Economic Botany*, **42**, 120–131.

Estoup A., Lombaert E., Marin J.M., Guillemaud T., Pudlo P., Robert C.P., & Cornuet J.M. (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources*, **12**, 846–855.

Evanno G., Regnaut S., & Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L., Dupanloup I., Huerta-Sánchez E., Sousa V.C., & Foll M. (2013) Robust demographic inference from genomic and SNP data. *PLoS genetics*, **9**, e1003905.

Excoffier L. & Lischer H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Fritz G. (1990) Multiple Pathways to Farming in Precontact North America. *Journal of World Prehistory*, **4**, 387–435.

Garza J.C. & Williamson E.G. (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Gillette C.P. (1912) *Diabrotica virgifera* Le Conte as a corn rootworm. *Journal of Economic Entomology*, **5**, 364–366.

Goudet J. (1995) FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. *Journal of Heredity*, **86**, 485–486.

Gray M.E., Sappington T.W., Miller N.J., Moeser J., & Bohn M.O. (2009) Adaptation and Invasiveness of Western Corn Rootworm: Intensifying Research on a Worsening Pest. *Annual Review of Entomology*, **54**, 303–321.

Hallatschek O. & Nelson D.R. (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Horn G.H. (1893) The Galerucini of Boreal America. *Transactions of the American Entomological Society*, **20**, 57–136.

Hufford M.B., Martínez-Meyer E., Gaut B.S., Eguiarte L.E., & Tenaillon M.I. (2012) Inferences from the Historical Distribution of Wild and Domesticated Maize Provide Ecological and Evolutionary Insight. *PLoS ONE*, **7**, .

Kim K.S., Ratcliffe S.T., French B.W., Liu L., & Sappington T.W. (2008) Utility of EST-Derived SSRs as population genetics markers in a beetle. *Journal of Heredity*, **99**, 112–124.

Kim K.S. & Sappington T.W. (2005) Genetic structuring of western corn rootworm (Coleoptera : Chrysomelidae) populations in the United States based on microsatellite loci analysis. *Environmental Entomology*, **34**, 494–503.

Krysan J.L. & Smith R.F. (1987) Systematics of the *virgifera* species group of *Diabrotica* (Coleoptera: Chrysomelidae: Galerucinae). *Entomography*, **5**, 375–484.

Langella O. (1999) Populations 1.2.32 (02/13/2011): a population genetic software. .

Lombaert E., Guillemaud T., Lundgren J., Koch R., Facon B., Grez A., Loomans A., Malausa T., Nedved O., Rhule E., Staverlokk A., Steenberg T., & Estoup A. (2014) Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the case of the Asian ladybird *Harmonia axyridis*. *Molecular ecology*, **23**, 5979–5997.

Marin J.-M., Raynal L., Pudlo P., Ribatet M., & Robert C.P. (2016) ABC random forests for Bayesian parameter inference. 1–16.

Meinke L.J., Sappington T.W., Onstad D.W., Guillemaud T., Miller N.J., Judith K., Nora L., Furlan L., Jozsef K., & Ferenc T. (2009) Western corn rootworm (Diabrotica virgifera virgifera LeConte) population dynamics. *Agricultural and Forest Entomology*, **11**, 29–46.

Merrill W.L., Hard R.J., Mabry J.B., Fritz G.J., Adams K.R., Roney J.R., & MacWilliams A.C. (2009) The diffusion of maize to the southwestern United States and its impact. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 21019–21026.

Metcalf R.L.. (1983) Implications and prognosis of resistance to insecticides. *Pest Resistance to Pesticides*, 703–733.

Metcalf R.L. (1986) The ecology of insecticides and the chemical control of insects. *Ecological theory and integrated pest management practice* (ed. by M. Kogan), pp. 251–297. New York.

Metcalf R.L. & Lampman R.L. (1989) The chemical ecology of Diabroticites and Cucurbitaceae. *Experientia*, **45**, 240–247.

Miller N., Estoup A., Toepfer S., Bourguet D., Lapchin L., Derridj S., Kim K.S., Reynaud P.,

Furlan L., & Guillemaud T. (2005) Multiple transatlantic introductions of the western corn rootworm. *Science*, **310**, 992.

Moeser J. & Hibbard B.E. (2005) A synopsis of the nutritional ecology of larvae and adults of Diabrotica virgifera virgifera (LeConte) in the new and old world - nouvelle cuisine for the invasive maize pest Diabrotica virgifera virgifera in Europe? *Western corn rootworm: ecology and management* (ed. by S. Vidal, U. Kuhlmann, and C.R. Edwards), pp. 41–65.

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

Oyediran I.O., Hibbard B.E., & Clark T.L. (2004) Prairie Grasses as Hosts of the Western Corn Rootworm (Coleoptera: Chrysomelidae). *Environmental Entomology*, **33**, 740–747.

Petit R.J., Mousadik A. El, & Pons O. (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.

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

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

R Development Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing. .

Rannala B. & Mountain J.L. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9197–9201.

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

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

Smith R.F. (1966) Distributional patterns of selected western north American insects: the distribution of Diabroticites in western north America. *Bulletin of the Entomology Society of America*, **12**, 108–110.

Tenaillon M.I. & Charcosset A. (2011) A European perspective on maize history. *Comptes Rendus - Biologies*, **334**, 221–228.

Vilà M., Basnou C., Gollasch S., Josefsson M., Pergl J., & Scalera R. (2009) One Hundred of the Most Invasive Alien Species in Europe. *Handbook of Alien Species in Europe* (ed. by DAISIE), pp. 265–268. Springer Netherlands,

Wegmann D., Leuenberger C., Neuenschwander S., & Excoffier L. (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *Bmc Bioinformatics*, **11**, 7.

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

Data accessibility

Data associated with this article (including microsatellite data file, ABC reference tables, input and script files for performing ABC simulations and analyses) are archived in Zenodo: http://doi.org/10.5281/zenodo.398844

Author contributions

EL and TG designed the study. TS managed the collection of samples. MC, NM and AB genotyped the samples. EL and TG analysed the data. EL, MC, NM, TS and TG wrote the paper. All authors have revised and approved the final manuscript.