MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/j.1365-294X.2012.05489.x

A Bayesian analysis of gene flow from crops to their wild relatives: cultivated (*Lactuca sativa* L.) and prickly lettuce (*L. serriola* L.) and the recent expansion of *L. serriola* in Europe

BRIGITTE UWIMANA,* LUIGI D'ANDREA,† FRANÇOIS FELBER,† DANNY A. P. HOOFTMAN,‡ HANS C. M. DEN NIJS,§ MARINUS J. M. SMULDERS,* RICHARD G. F. VISSER* and CLEMENS C. M. VAN DE WIEL*

*Wageningen UR Plant Breeding, Wageningen, The Netherlands, †Laboratoire de Botanique évolutive, Université de Neuchâtel, Neuchâtel, Switzerland, ‡Centre for Ecology & Hydrology, Wallingford, UK, §IBED, Universiteit van Amsterdam, Amsterdam, The Netherlands

Abstract

Interspecific gene flow can lead to the formation of hybrid populations that have a competitive advantage over the parental populations, even for hybrids from a cross between crops and wild relatives. Wild prickly lettuce (*Lactuca serriola*) has recently expanded in Europe and hybridization with the related crop species (cultivated lettuce, *L. sativa*) has been hypothesized as one of the mechanisms behind this expansion. In a basically selfing species, such as lettuce, assessing hybridization in natural populations may not be straightforward. Therefore, we analysed a uniquely large data set of plants genotyped with SSR (simple sequence repeat) markers with two programs for Bayesian population genetic analysis, STRUCTURE and NewHybrids. The data set comprised 7738 plants, including a complete genebank collection, which provided a wide coverage of cultivated germplasm and a fair coverage of wild accessions, and a set of wild populations recently sampled across Europe. STRUCTURE analysis inferred the occurrence of hybrids at a level of 7% across Europe. NewHybrids indicated these hybrids to be advanced selfed generations of a hybridization event or of one backcross after such an event, which is according to expectations for a basically selfing species. These advanced selfed generations could not be detected effectively with crop-specific alleles. In the northern part of Europe, where the expansion of *L. serriola* took place, the fewest putative hybrids were found. Therefore, we conclude that other mechanisms than crop/wild gene flow, such as an increase in disturbed habitats and/or climate warming, are more likely explanations for this expansion.

Keywords: gene flow, hybrid identification, molecular markers, population genetics

Received 10 May 2011; revision received 14 December 2011; accepted 20 December 2011

Introduction

Gene flow through hybridization is a common phenomenon among closely related plant species. Recent studies have shown it to be more frequent between crop species and their wild relatives than assumed based on the supposition that domestication traits are probably to reduce fitness under natural conditions (Ellstrand 2003). For instance, gene flow was reported to occur between 12 of the 13 most important food crops and their respective wild relatives (Ellstrand *et al.* 1999). With the large-scale cultivation of genetically modified cultivars, gene flow from crops to their wild relatives has attracted public interest and concern and has initiated research on gene escape and introgression in the framework

Correspondence: Brigitte Uwimana, Fax: +31 317 483457;
E-mail: brigumbu@gmail.com or clemens.vandewiel@wur.nl

of environmental risk assessment of transgenic plants (Pilson & Prendeville 2004; Snow *et al.* 2005; Chapman & Burke 2006; Warwick *et al.* 2008, 2009).

Gene flow may have evolutionary impact, especially when particular genes from crops would increase the fitness, and thus possibly the invasiveness of weeds, for instance, by increasing their adaptability to various climatic or environmental conditions (Langevin *et al.* 1990; Ellstrand & Schierenbeck 2000; Magnussen & Hauser 2007). In this regard, an invasive trend was reported for wild (weedy) prickly lettuce (*Lactuca serriola*), the closest wild relative of cultivated lettuce (*L. sativa*) in many Mediterranean, Central and Western European countries (Frietema de Vries *et al.* 1994; Lebeda *et al.* 2004b). Hooftman *et al.* (2006) reported a sweeping spread of *L. serriola* also in the Netherlands since 1980 and lists four possible reasons for this recent invasiveness: (i) a change in environment owing to global warming; (ii) increased landscape disturbance by human activities, which produces more suitable habitat; (iii) micro-evolution of the species towards extended adaptability; and (iv) hybridization between *L. serriola* and cultivated *L. sativa*. The latter reason would be a direct consequence of gene flow through interspecific hybridization, leading to the transfer and introgression of genes from the crop to the wild lettuce that confer increased fitness to the resulting crop/wild hybrids.

Cultivated *L. sativa* was most likely domesticated from ancient populations of *L. serriola* (De Vries 1997). Whereas De Vries & Van Raamsdonk (1994) regarded them as separate species based on multivariate numerical morphological analysis, Frietema de Vries *et al.* (1994) postulated them to be conspecific based on a morphological analysis of a large array of crop types and wild accessions. Koopman *et al.* (2001) supported the view of Frietema de Vries *et al.* (1994) on the basis of their analysis of species relationships in the genus *Lactuca*, using AFLP markers. De Vries (1990) showed that *L. sativa* can be crossed with *L. serriola* to form viable and fertile hybrids. Even though *L. sativa* and *L. serriola* are basically self-pollinators, Thompson *et al.* (1958) reported an out-crossing rate of 1–5% among *L. sativa* varieties, and D'Andrea *et al.* (2008) reported an interspecific hybridization rate of up to 2.5% between the two species. Hooftman *et al.* (2005, 2009) studied the performance of the hybrids resulting from manual crosses between *L. serriola* and *L. sativa*, and Hooftman *et al.* (2007) modelled the long-term consequences of hybridization between the two species. These two studies established that hybrids between *L. sativa* and *L. serriola* are viable, fertile and that the hybrid offspring may even be fitter than the wild parent. Based on single fitness components, hardly any significant differences were detected between prickly lettuce and the hybrid plants, and backcrossed hybrids were morphologically indistinguishable from their wild parent (*L. serriola*) (Hooftman *et al.* 2005). Hence, the fact that only very few occurrences of crop/wild hybrid lettuce in the field have been reported (*cf.* Frietema de Vries *et al.* 1994) is not necessarily proof of a lack of occurrence as it may, at least in part, be due to problems in recognizing putative hybrids.

In this study, we aimed to quantify the spontaneous occurrence of gene flow between cultivated and wild lettuce in Europe. A number of methods can be used for the identification of hybrid plants in natural populations of wild relatives of crop species, including screening based on phenotypic traits (Ureta *et al.* 2008), tracking crop-specific markers (Westman *et al.* 2001; Morrell *et al.* 2005; Scurrah *et al.* 2008) and, in case of GM crops, tracking the transgene itself (Warwick *et al.* 2008). These methods do not work well in all cases. As already indicated, for this study, the use of morphological traits would be difficult because the hybrids resulting from crosses between *L. serriola* and *L. sativa* often look morphologically like *L. serriola* (Hooftman *et al.* 2005). The use of a transgene as a marker is also not applicable outside of contained conditions because, there is not (yet) any transgenic lettuce cultivar allowed for commercial cultivation. The 'crop-specific' allele approach scans each locus for alleles with differences in occurrence between crop and wild relatives. Alleles far more frequent in crops are then used as indications for introgression from the crop when found in wild plants. This method has been regularly used to trace hybridization between crops and wild relatives, but suffers from two problems: (i) the definition of 'crop-specific' alleles; and (ii) how to distinguish their occurrence as a result of recent introgression from one as a result of a more ancient common ancestry (e.g. Van de Wiel *et al.* 2005). Thus, for this study, we used two Bayesian posterior probability-based methods, one as implemented in the software package STRUCTURE (Pritchard *et al.* 2000) and the other as implemented in NewHybrids (Anderson & Thompson 2002), to analyse two large data sets of lettuce samples, one from the genebank collection of crop (*L. sativa*) and wild lettuce (*L. serriola*), and the other set comprising of *L. serriola* samples collected across Europe in the period of 2002–2005. Together, these data sets constitute an exceptionally broad set for a study of gene flow between crops and hybrids. When Smulders *et al.* (2008) compared STRUCTURE with NewHybrids to detect hybridization with cultivated poplar hybrids in offspring of wild *Populus nigra* trees, they found that NewHybrids was more informative on the degree of hybridization. As lettuce is highly selfing, we anticipated that putative hybrids would have a high likelihood of being advanced selfed generations. We

therefore applied STRUCTURE to identify potential hybrid plants and NewHybrids to infer the number of selfings or backcrossings after an initial hybridization event between *L. serriola* and *L. sativa*. STRUCTURE results were checked using the program InStruct (Gao *et al.* 2007), which takes into account the divergence from the Hardy–Weinberg equilibrium owing to self-fertilization. We compared the STRUCTURE results with a crop-specific allele approach to assess to what extent the latter method still has its value with regard to its relative easy implementation for detecting gene flow between crop and wild relatives.

Materials and methods

Plant material

We studied the crop/weed complex of cultivated lettuce (*L. sativa*) and wild prickly lettuce (*L. serriola*). *L. serriola* is a weed plant, which thrives in anthropogenically disturbed areas (Lebeda *et al.* 2001), whereas *L. sativa* is a vegetable crop species. *L. serriola* is the closest relative of *L. sativa*, with the latter considered as the direct descendent of *L. serriola* (Kesseli *et al.* 1991; Frietema de Vries *et al.* 1994). The two species have the same number of chromosomes ($2n = 18$) and their close relatedness has led to some studies classifying them as conspecific (Koopman *et al.* 1993, 2001). They are readily crossable without any known barrier and their hybrids are viable and fertile (De Vries 1990). *L. serriola* is mostly distinct from *L. sativa* based on their morphological traits (De Vries & Van Raamsdonk 1994), but their hybrids, especially those resulting from backcrosses to *L. serriola*, are generally not distinguishable from the latter (Hooftman *et al.* 2005). The two species grow often in sympatry in Southern Europe where lettuce seed production is carried out in open fields, and in home gardens where cultivated lettuce is often left to flower, leading to opportunities for interspecific hybridization (D'Andrea *et al.* 2009).

We used plant material originating from two sources: the lettuce collection from the Centre for Genetic Resources, the Netherlands (CGN) and a recent collection of *L. serriola* from across Europe. CGN hosts the largest lettuce germplasm collection worldwide (http://www.cgn.wur.nl/UK/CGN+Plant+Genetic+Resources/Collections/Leafy+vegetables/Lettuce/), which has a comprehensive representation of genetic variation in cultivated lettuce, supplemented with a fair representation of wild relatives, particularly of *L. serriola* (Van de Wiel *et al.* 2010). This collection comprises lettuce accessions collected since 1940, with some over-representation of germplasm from Europe and the Middle East (Lebeda *et al.* 2004a). *L. serriola* accessions of this collection used in this study are designated as 'CGN *L. serriola*' and *L. sativa* accessions are designated '*L. sativa*'.

The recent *L. serriola* collection was sampled from 2002 to 2005 within the EU project 'Analysis of gene flow from crop to wild forms in lettuce and chicory and its population-ecological consequences in the context of GM-crop biosafety' (ANGEL, QLK3-CT-2001-01657, http://www.plant.wageningen-ur.nl/projects/angel/) (Van de Wiel *et al.* 2003). These *L. serriola* individuals are designated as 'ANGEL *L. serriola*'. They were collected in 17 European countries (Austria, Bulgaria, Croatia, Czech Republic, Denmark, France, Germany, Greece, Hungary, Italy, Luxembourg, the Netherlands, Poland, Portugal, Slovakia, Spain, Sweden and Switzerland) from ruderal sites such as roadsides, along railways, vicinities of riverbanks, crop fields and construction sites, and in vegetable kitchen gardens. The kitchen garden locations included 'amateur gardens' in the Netherlands where *L. sativa* plants are more often left to flower than in professional cultivations and thus hybridization may have a good likelihood of occurrence (Hooftman *et al.* 2005). The particular advantage of the ANGEL and CGN data sets is that CGN has a comprehensive representation of germplasm cultivated worldwide as well as wild material collected from a large part of the areas where *L. serriola* occurs, and the ANGEL samples added more details and density on recent wild populations across the European continent.

Genotyping

Ten SSR markers, originally described by Van de Wiel *et al.* (1999), were used to genotype the individual plants (Table 1). The genetic positions for these ten markers have been determined on the lettuce genetic map (Truco *et al.* 2007). Eight marker loci were located on 8 different linkage groups, and two loci were located on the ninth chromosome but with a distance of 86 cM. Thus, the loci were considered as genetically unlinked. The CGN samples were genotyped under the EU project 'Molecular markers for genebanks: Application of marker technology for the improvement of ex situ germplasm conservation methodology' (PL96.2062) using a gel-based ABI PRISM® 377 DNA Sequencer (Applied Biosystems) (Van Hintum 2003; Van Treuren *et al.* 2008). The ANGEL samples were genotyped using a capillary-based ABI PRISM® 3700 DNA Sequencer (Applied Biosystems). The two genotyping methods were checked for consistency and concordance by genotyping a random sample of CGN gel-scored accessions using the capillary method and using three standard samples across all runs (Van Treuren *et al.* 2008). The random sample consisted of five individuals for each

Table 1 Description of the SSR markers used for genotyping

Locus*	Repeat motif	Linkage group	Observed number of alleles	Effective number of alleles	F_{IS}†,‡		
					ANGEL *Lactuca serriola*	CGN *L. serriola*	*L. sativa*
LsA001	$(GA)_{44}(GT)_{11}$	1	51	12.7	0.98	0.98	0.99
LsA004	$(GA)_{19}(GT)_7(GAGT)_4(GA)_{10}(GAGT)_2$ $(GA)_{21}(GT)_{12}$	6	27	8.6	0.98	0.99	0.99
LsB101	$(GT)_{12}(AT)_5(GT)_{17}$	2	31	7.7	0.98	0.99	0.99
LsB104	$(GA)_5(GT)_7TATT(GT)_{12}(T)_4(GT)_8(GA)_{11}$	4	36	7.7	0.98	0.99	0.99
LsD103	$(TCT)_{17}$	9	14	5.6	0.98	0.99	0.99
LsD106	$(TCT)_{17}(T)_5(TCT)_2$	5	16	5.6	0.99	0.98	0.98
D108	$(TCT)_{35}$	4	48	11.7	0.99	0.98	0.98
D109	$(TCT)_{22}$	8	34	14.4	0.97	0.98	0.99
LsE003	$(TGT)_{24}(TA)(TGT)_{10}(TAT)_2(TGT)_3$	7	24	4.1	0.98	0.98	0.98
E011	$(TGT)_{26}$	3	24	4.6	0.98	0.98	0.99
All					0.98	0.98	0.99

*Originally described by Van de Wiel *et al.* (1999).
†Calculated with FSTAT version 2.9.3.2. (Goudet 1995 for version 1.2).
‡Reduction in heterozygosity because of inbreeding.

CGN *L. serriola* accession and two individuals for each *L. sativa* accession as the crop accessions were expected to be more uniform than the wild accessions. Each ANGEL *L. serriola* collection site was represented by 30 individuals, which were all genotyped, and each site was considered as an ecological population [see Online Supplementary material I (Appendix S1, Supporting information) for details on origin of samples]. The loci amplified well in all the three data sets except for locus *LsD103* which had a poorer amplification in both ANGEL and CGN *L. serriola* samples than in *L. sativa* samples. As expected for predominantly selfing species, inbreeding caused great reduction in heterozygosity (F_{IS} ranging from 0.97 to 0.99, Table 1). CGN and ANGEL *L. serriola* data sets were not genetically different (F_{ST} = 0.02), whereas *L. serriola* and *L. sativa* samples showed a moderate level of differentiation with F_{ST} between ANGEL *L. serriola* and *L. sativa* = 0.24 and F_{ST} between CGN *L. serriola* and *L. sativa* = 0.20 (F_{ST} estimated using FSTAT v. 2.9.3.2.; Goudet 1995).

After genotyping, individuals with more than 50% missing data were removed, together with CGN *L. serriola* samples whose country of origin was not recorded in the CGN passport data. In total, 7738 individuals remained: 2456 ANGEL *L. serriola* samples, 2462 CGN *L. serriola* samples and 2820 *L. sativa* samples. The ten markers used for genotyping resulted in 14–54 alleles per locus. The effective number of alleles per locus, calculated as $1/\sum P_i^2$, with P_i = allele frequency (Storme *et al.* 2004), ranged from 4 to 14 (Table 1), as many alleles were rare in both wild and cultivated lettuce.

Data analysis

Determination of population structure using STRUCTURE and InStruct. Analysis for population structure was performed on the 7738 individuals using STRUCTURE (Pritchard *et al.* 2000) version 2.2 (Falush *et al.* 2007). It uses a model-based Bayesian method to cluster the plant samples in a number of clusters (K) based on their genotypes. The ancestry model was set to admixture with correlated allele frequencies and lambda 1.0. No prior population information was used in the analysis. After a number of combinations of burn-in and Markov Chain Monte Carlo (MCMC) runs, the final number of runs was chosen so that the differences in likelihood for each run [ln$P(D)$] between different K's was larger than the variability within runs of the same K. The length of the burn-in period was set to 80 000 with 150 000 MCMC replication runs after burn-in. To identify potential crop/wild hybrids, STRUCTURE should correctly differentiate *L. serriola* from *L. sativa* based on the 10 SSRs. To avoid any bias, we did not impose two clusters (K = 2) on the program, but we let it run from K = 1 to 35 to check whether it would differentiate the two species or come to alternative subdivisions. The median of ln$P(D)$ for six iterations of each K was considered (Saisho & Purugganan 2007), and the optimum number of clusters was determined by looking at the value of K with the highest likelihood (Pritchard *et al.* 2000) and the K value where the maximum number of information was gained in the analysis, that is where the ln$P(D)$ value increased most from one K to the next (Evanno *et al.* 2005). To differentiate nonadmixed and

admixed (potentially hybrid) plants, we used a threshold posterior probability (*Q*-value) of 0.90. Plants with *Q*-value equal to or greater than 0.90 were considered as nonadmixed; and those with *Q*-value smaller than 0.90 were considered as admixed or potential hybrids (Vähä & Primmer 2006; Burgarella *et al*. 2009).

The data were also analysed with InStruct (Gao *et al*. 2007), a Bayesian-based program similar to STRUCTURE but specifically written for selfing species to account for divergence from the Hardy–Weinberg equilibrium owing to self-fertilization. The length of the burn-in period was set to 100 000 and 200 000 MCMC runs after burn-in with $K = 1–10$ and five iterations for each *K*. The optimum number of clusters and the classification of individuals as nonadmixed or hybrid were carried out as described earlier for the STRUCTURE results.

Inference of the hybrid generations using NewHybrids. NewHybrids (Anderson & Thompson 2002) is Bayesian model-based software to calculate the posterior probability that each plant belongs to a certain category of parents or hybrids based on the genotypic information of the plants. We used NewHybrids version 1.1 to infer the generations of the hybrid plants identified by STRUCTURE. Because of the limited capacity of the software, only a subset of the samples analysed with STRUCTURE was used with NewHybrids, namely all the hybrids identified by STRUCTURE and a randomly chosen set of nonadmixed plants from the two *L. serriola* data sets and *L. sativa* plants. These were 706 ANGEL *L. serriola*, 617 CGN *L. serriola* and 677 *L. sativa* individuals, totalling 2000 individuals. *L. serriola* and *L. sativa* being basically self-pollinating species, hybrid plants were expected to belong to advanced selfing generations after either one cross between the two species or one backcross to any of the two parents. It is impossible to reliably distinguish between various advanced generations based on the 10 markers used in this study, owing to little change in heterozygosity from one advanced generation to another. Therefore, we only considered the following categories in this analysis: nonadmixed *L. serriola* and nonadmixed *L. sativa* as the parents of the hybrids (Parent 0 and Parent 1, respectively), early generations of selfing after one cross between the two parents (F1 and F2), and early generations of selfing after one backcross to either of the two parents (BC1 and BC1S1). Advanced generations were represented by three categories, namely F7 (selfing without backcrossing), 0-BC1S7 (selfing after one backcross to *L. serriola*) and 1-BC1S7 (selfing after one backcross to *L. sativa*; Fig. 1). For early generations, a high frequency of heterozygosity was expected (100% for F1 and 50% for F2 and BC1), whereas with advanced generations heterozygosity was expected to be close to zero (1.6% for F7 and 0.4% for BC1S7). We ran NewHybrids using the uniform prior for both θ and the mixing proportion π and the program was left to run for 900 000 sweeps after burn-in. Because NewHybrids is less sensitive than STRUCTURE in differentiating between nonadmixed and admixed individuals (Vähä & Primmer 2006), a threshold posterior probability (*P* value) of 0.70 was used to categorize an individual as belonging to a specific group.

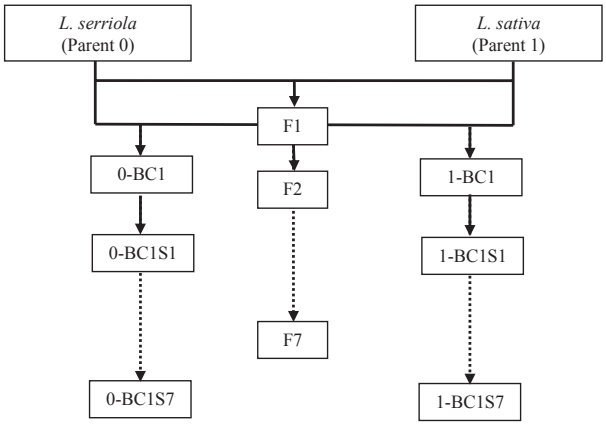

Fig. 1 Hybrid classes used in NewHybrids: the hybrids were categorized into early and advanced generations of selfing after one cross between the two parents and early and advanced generations of selfing after one backcross to either of the two parents, with the advanced generations representing the inbred generations. Advanced inbred generations cannot be differentiated owing to limited change in heterozygosity from one generation to the next.

Is crop/wild hybridization the cause of the spread of L. serriola in Europe?. To test whether crop/wild hybridization is the cause of the northward spread of *L. serriola* in Europe, a Pearson's Chi-square test of independence was run on the most recent collection of *L. serriola* in Europe (ANGEL data set), testing whether the occurrence of the hybrids depended on the region where the samples were collected. The origin of the samples was categorized in two groups based on the plant geographical regions of Europe (Schaminée *et al*. 1992; Frietema de Vries *et al*. 1994). The Southern region was represented by Portugal, Spain, south of France (below 45° of latitude), Italy, Switzerland, Hungary and Austria. The Northern region consisted of the remaining part of France, Luxembourg, the Netherlands, Germany, Czech Republic and Denmark. If crop/wild hybridization is responsible for the northward spread of *L. serriola*, we expect the Chi-square test to show a dependence between the occurrence of hybrids and the

region (North and South) where the samples were collected and a bigger proportion of hybrids compared with nonhybrids should be found in the Northern region.

Results

Distinction between wild L. serriola *and cultivated* L. sativa *by STRUCTURE*

With the STRUCTURE analysis, ln*P*(*D*) gradually increased and no clear peak was reached up to $K = 35$ (Fig. 2A). Therefore, the choice of the number of clusters, *K*, was not based on the highest ln*P*(*D*) value but on the value of *K* where the maximum information was gained in the analysis based on difference in ln*P*(*D*) from one *K* to the next (Delta *K*, Evanno *et al.* 2005). Maximum information was gained from $K = 1$ to 2 (Fig. 2B). At $K = 2$, STRUCTURE well-differentiated *L. sativa* from *L. serriola*. The plants with high posterior probability (>0.90) to one of the two groups coincided with *L. serriola* and the plants with high posterior probability to the other group coincided with *L. sativa*. 'Admixed' plants with intermediate probabilities to both groups were considered as potential hybrids (Fig. 3). To check whether the differentiation achieved by STRUCTURE between *L. serriola* and *L. sativa* was consistent, we checked *K* values larger than 2. At $K = 3$, *L. serriola* remained distinct from *L. sativa*, with the third cluster arising by a split of the *L. sativa* cluster. At $K = 4$, *L. serriola* samples were again clearly distinct from *L. sativa* samples, with both *L. sativa* and *L. serriola* being split into two clusters (Online Supplementary material II, Fig. S1, Supporting information). However, these clusters did not coincide with any recognizable biological or geographical group.

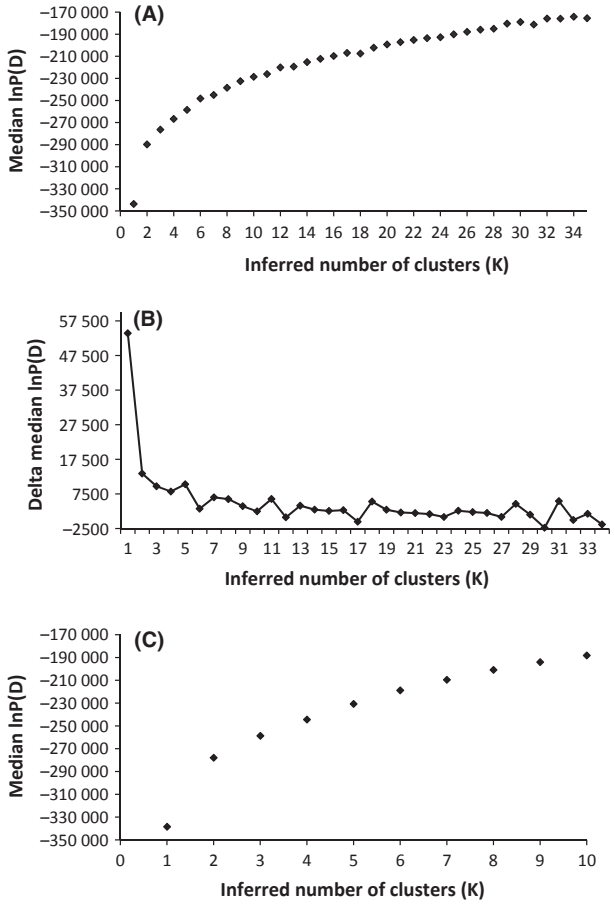

Fig. 2 Analysis with STRUCTURE and InStruct. (A) STRUCTURE ln*P*(*D*) median as a function of the number of inferred clusters, *K*, up to $K = 35$. (B) Delta ln*P*(*D*) for subsequent *K*s from STRUCTURE: maximum information was gained between $K = 1$ and $K = 2$. (C) InStruct ln*P*(*D*) median as a function of the number of inferred clusters *K*, up to $K = 10$.

Occurrence of 'admixed' (hybrid) plants

At $K = 2$, potential hybrids had intermediate probabilities to the two clusters that coincided with *L. serriola* and *L. sativa*. Because the two clusters mirrored each other (as a *Q*-value of 0.90 for one cluster is equivalent to 0.10 for the other), samples with *Q*-values smaller than 0.10 were regarded as nonadmixed *L. sativa* and samples with *Q*-values >0.90 as nonadmixed *L. serriola* in the remainder of this study. Ninety-three per cent of ANGEL *L. serriola* and 87% of CGN *L. serriola* individuals clustered in these groups, resulting in 7% potential hybrids (181 plants) among the ANGEL *L. serriola* individuals and 13% (312 plants) in the CGN *L. serriola* data set (Fig. 4A). The CGN data set not only had a greater proportion of admixed individuals,

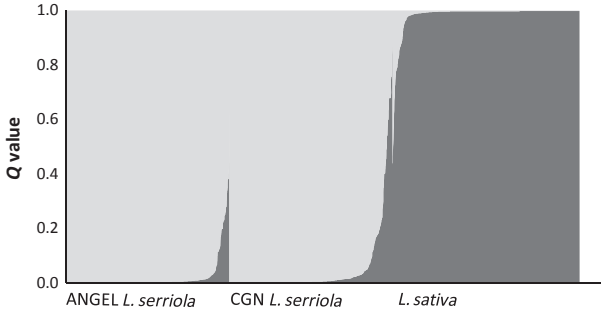

Fig. 3 *Q* graph of all the samples at $K = 2$. The *y*-axis represents the *Q*-values and the *x*-axis represents the plants: each plant is given a membership proportion into two groups (grey and white), corresponding to *Lactuca serriola* and *L. sativa*, respectively. Putative hybrids are discernible as partly belonging to both parental groups, *L. sativa* (right) and the combined data sets of *L. serriola* (left).

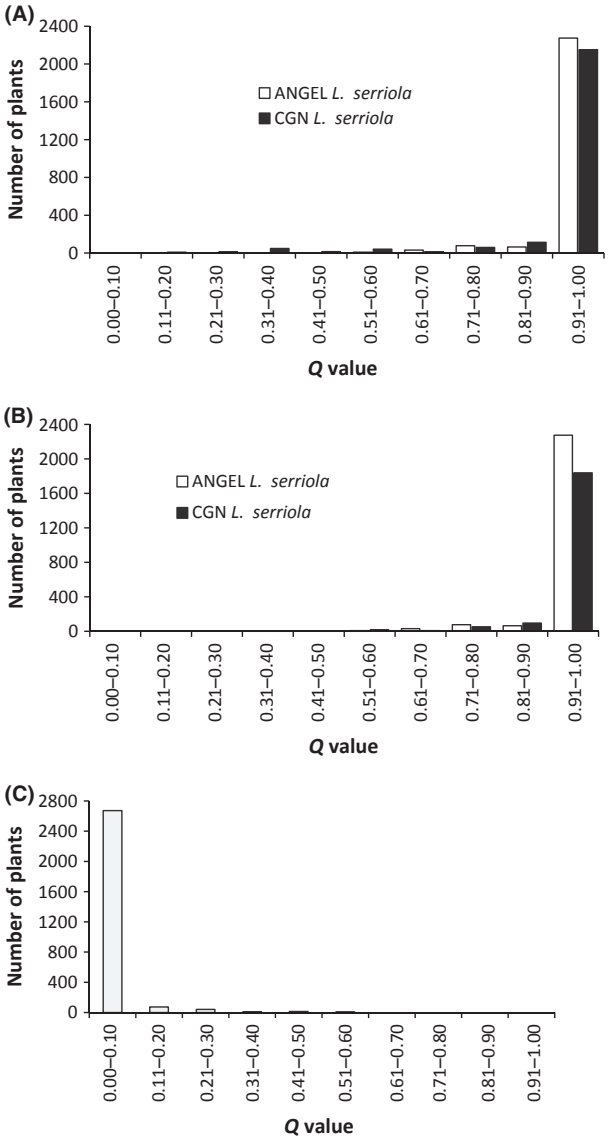

Fig. 4 Number of individuals for each *Q*-value range (grouped using a range of 0.10) at $K = 2$. (A) ANGEL and CGN *Lactuca serriola* data sets; (B) ANGEL and CGN *L. serriola* without the *L. serriola*-CGN plants with uncertain origin data; (C) *L. sativa* (cultivated lettuce) data set.

but its putative hybrids also showed a more extended *Q*-value range (0.13–0.90) than the putative hybrids in the ANGEL data set (0.32–0.90). On closer scrutiny of the CGN passport data the origin of the CGN accessions proved to be the major cause of the differences between the two data sets. Even though we had removed the samples whose origin was completely unknown before performing the analysis, CGN *L. serriola* hybrids with the lowest *Q*-values (0.13–0.35, so most '*L. sativa*'-like) were not obtained directly from their original habitats, but through research institutions or botanical gardens (Lebeda *et al.* 2004a). Such accessions have been shown to deviate genetically from material with an established origin in the same region, and some were even genetically identical to accessions from botanical gardens in other, distant, countries (Van de Wiel *et al.* 2010). By excluding all accessions without clearly established origins, the CGN *L. serriola* data set became more similar to the ANGEL data set: the lowest *Q*-value for CGN *L. serriola* hybrids increased to 0.40, and the proportion of hybrids dropped to 9% (Fig. 4B). Among *L. sativa* 5% (147 plants, Fig. 4C) clustered as putative hybrids, with *Q*-values ranging from 0.11 (close to nonadmixed *L. sativa*) to 0.57.

InStruct gave similar results as STRUCTURE. Ln$P(D)$ did not show any peak up to $K = 10$ and the maximum information was gained from $K = 1$ to 2, making $K = 2$ the optimum number of clusters (Fig. 2C). At a threshold posterior probability of 0.90, the two programs classified 98% of the plants in the same categories (Table 2). The two per cent dissimilarity arose from InStruct identifying more *L. serriola* hybrids (0.16% among ANGEL and 3.25% among CGN *L. serriola*) and fewer *L. sativa* hybrids (0.39%) than STRUCTURE. These dissimilarities between InStruct and STRUCTURE were because of small differences in *Q*-values which ranged from 0.01 to 0.08. The results by STRUCTURE were more conservative, and hence, we used them for further analysis (Arrigo *et al.* 2011).

In Europe, the putative *L. serriola* hybrids were more frequent in the South (Fig. 5). In the ANGEL data set, most of the putative hybrids were found in Spain, Portugal, Italy and southern France: 141 of the 181 ANGEL potential hybrids came from this region. *Q*-values in the region were as low as 0.32. In the more northerly country of the Netherlands, only 10 of 152 samples (6%) were putative hybrids (*Q*-values 0.70–0.90). The 28 plants collected in the direct vicinity of amateur gardens did not indicate any increased likelihood of hybridization, as only two of these plants were identified as hybrids (with $Q = 0.70$). These represented 7% of the hybrid occurrence, which was similar to that of the randomly sampled populations. Taken together, the proportion of hybrids compared to nonadmixed individuals was 10% in the Southern region, while it was 2% in the Northern region. A Chi-square test of independence showed that the occurrence of the hybrids differed between these regions ($P < 0.001$, Table 3).

In the CGN *L. serriola* data set, there was no difference in hybrid occurrence between Europe and the Middle East (9% and 10%, respectively). These figures could be biased due to the over-representation of accessions from Europe in the genebank (Lebeda *et al.*

Table 2 Comparison between the classification of hybrids by STUCTURE and InStruct: the numbers in the table are the percentages of the classification by STRUCTURE. The differentially classified individuals are underlined. The two programs categorize 98% of all the data in the same classes of nonadmixed and potential hybrids

	InStruct						
	ANGEL *Lactuca serriola*		CGN *L. serriola*		*L. sativa*		
STRUCTURE	Nonadmixed	Hybrids	Nonadmixed	Hybrids	Nonadmixed	Hybrids	Total STRUCTURE (number of individuals)
ANGEL *L. serriola*							
Nonadmixed*	99.78	0.22					2275
Hybrids	0.55	99.45					181
CGN *L. serriola*							
Nonadmixed			95.77	4.23			2150
Hybrids			3.53	96.47			312
L. sativa							
Nonadmixed					99.29	0.71	2673
Hybrids					20.41	79.59	147
Total Instruct (number of individuals)	2271	185	2070	392	2684	136	7738

*Nonadmixed: $Q > 0.90$; hybrids: $Q \leq 0.90$.

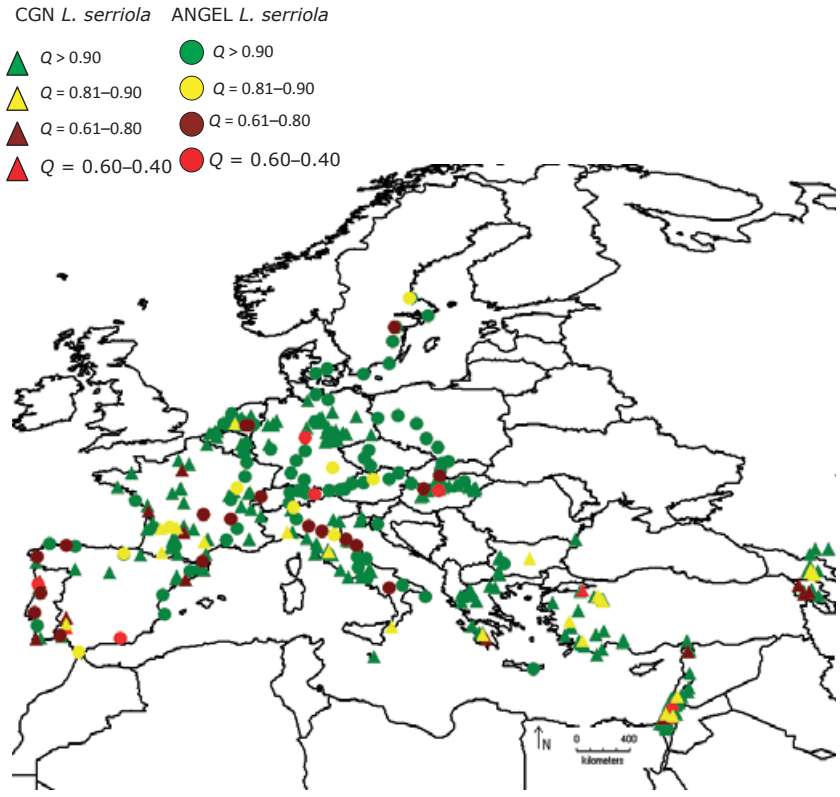

Fig. 5 Geographical origin of nonadmixed (green triangles and circles) and putative hybrid *Lactuca serriola* plants (other coloured triangles and circles, see legend) as identified by STRUCTURE in ANGEL and CGN data sets. At each location, the lowest *Q*-value is represented.

2004a). Outside of Europe, most of the *L. serriola* individuals from the wild habitats were collected from Israel and Turkey (89% of all Middle East individuals, Fig. 5). The occurrence of hybrids in these two countries was 9%, which reflected the same pattern of hybrid occurrence as in the whole region.

Table 3 Contingency table for the Chi-square test of independence between the occurrence of hybrids among ANGEL *Lactuca serriola* individuals and the region where the samples were collected

	Hybrids	Nonadmixed	Total
South			
Count	151	1135	1756
Expected	96	1190	
North			
Count	30	1119	700
Expected	85	1064	
Total	181	2277	2456
Chi-square test of independence			73.52 ($P < 0.001$)

Inferences about likely generations of putative hybrid plants by NewHybrids

NewHybrids classified nonadmixed plants as *L. sativa* and *L. serriola*. Admixed plants belonged to advanced selfing generations after one hybridization event between *L. serriola* and *L. sativa* (represented by F7) and to advanced selfed generations after one backcross to either *L. serriola* or *L. sativa* (represented by 0_BC1S7 and 1_BC1S7, Fig. 6). Other classes were not represented at all or were represented by probabilities smaller than 0.006. At a threshold *P* value of 0.70, NewHybrids classified 1265 of the 2000 plants in one of the five categories. The remaining 735 had probabilities divided between two or three categories (Table 4).

We compared the NewHybrids and STRUCTURE results (Table 4). NewHybrids recognized as hybrids all 181 ANGEL *L. serriola* plants identified as potential hybrids by STRUCTURE, and 97% of the 312 CGN *L. serriola* plants that STRUCTURE identified as potential hybrids. For *L. sativa*, NewHybrids recognized as hybrids 99% of the plants that STRUCTURE identified as hybrids. Conversely, NewHybrids also classified many of the STRUCTURE nonadmixed plants as hybrids. Ten per cent of the ANGEL *L. serriola* plants identified as nonadmixed by STRUCTURE were recognized as either hybrid or *L. serriola* (with higher probabilities for the *L. serriola* class, $P > 0.45$), and 8% were classified as hybrids by NewHybrids. Of the CGN *L. serriola* plants identified by STUCTURE as nonadmixed, NewHybrids classified 21% as undecided between nonadmixed *L. serriola* and hybrids, and 34% as hybrids. Of the *L. sativa* plants identified by STRUCTURE as nonadmixed, NewHybrids classified 10% as either nonadmixed or hybrids, and 13% as hybrids.

Comparison between STRUCTURE and crop-specific allele method in identifying hybrids

To assign alleles as 'crop-specific', we used the following criteria: (i) the frequency of the allele among *L. sativa* individuals is at least an order of two magnitudes higher than its frequency in the *L. serriola* data sets (restricted to accessions with confirmed origin data for the CGN data set) and (ii) to attain a fair level of representativeness, the putative 'crop-specific' alleles should occur in more than 10% of the accessions of the *L. sativa* data set. Only six alleles from five loci conformed to our criteria of 'crop-specificity' (LsA001-187, LsD103-263, LsD103-266, LsD106-191, LsE003-206 and E011-251); two additional alleles (D109-251 and E011-254) conformed in the ANGEL *L. serriola* set only (Table S1, Supporting information). For the CGN *L. serriola* set, both sets of alleles matched these criteria only when accessions from Europe were exclusively taken into account. The considerably higher frequency of the D109-251 and E011-254 alleles in accessions from the Middle East and Central Asia could be related to this area being the most likely centre of origin of cultivated lettuce.

Table 5 shows a comparison of hybrid identification results from these crop-specific alleles with those from

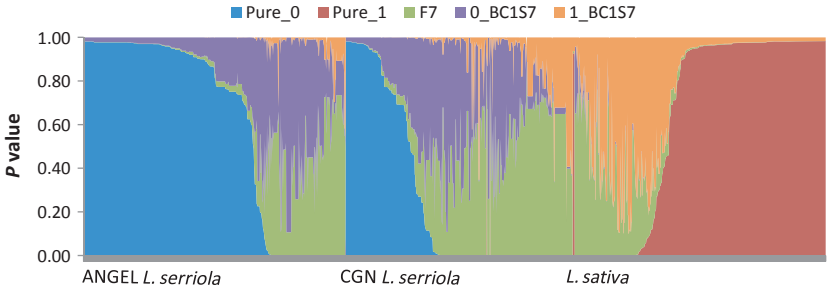

Fig. 6 Average category probabilities of all analysed plants by NewHybrids. Each class is represented by a colour, on the *x*-axis are individual plants and the *y*-axis represents the probability values, which add up to 1. The plants categorize as nonadmixed *Lactuca serriola* and *L. sativa* and advanced hybrid generations. The remaining classes included in the analysis were not represented at all or were represented by very small probability values ($P < 0.006$).

Table 4 Comparison between NewHybrids and STRUCTURE results: the numbers in the table are percentages of the classification by STRUCTURE. The differentially classified individuals are underlined. NewHybrids classified the hybrids as 7th generation of selfing after the initial cross between *Lactuca serriola* and *L. sativa* (F7) or the 7th generation of selfing after one back-cross to either *L. serriola* (0-BC1S7) or *L. sativa* (1-BC1S7) (see Fig. 1 for an overview of classes). The '7th generation' represents advanced selfing generations. The differentially classified individuals by STRUCTURE and NewHybrids are underlined

	NewHybrids											
STRUCTURE	*L. serriola* (Parent 0)	*L. sativa* (Parent 1)	F7	0-BC1S7	1-BC1S7	F7 or 0-BC1S7	*L. serriola* or F7or 0-BC1S7	*L. serriola* or 0-BC1S7	F7 or 1-BC1S7	*L. sativa* or 1-BC1S7	*L. sativa* or F7 or 1-BC1S7	Total STRUCTURE (number of individuals)
ANGEL *L. serriola*												
Nonadmixed	82.82			1.34		6.30	3.05	6.49				524
Hybrids			19.89	22.10		53.59			4.42			181
CGN *L. serriola*												
Non-admixed	44.26			2.30		29.18	8.20	13.11	2.95			305
Hybrids	0.64	2.24	12.82	10.58	0.64	41.35		2.56	29.17			312
L. sativa												
Nonadmixed		77.55	0.75	-	4.72	-	-	-	7.36	7.92	1.70	530
Hybrids		0.68	19.73		35.37	4.08			40.14			147
Total NewHybrids (number of individuals)	571	419	109	87	79	355	41	82	206	42		2000

Table 5 Frequency of crop-specific alleles* among *Lactuca serriola* data sets categorized as potential hybrids and nonadmixed using STRUCTURE

STRUCTURE groups	ANGEL *L. serriola*	CGN *L. serriola* from Europe	CGN *L. serriola* from outside of Europe
Hybrids ($Q \leq 0.90$)			
Frequency STRUCTURE	181	84	93
Frequency of plants containing at least one crop-specific allele	16	2	14
Nonadmixed ($Q > 0.90$)			
Frequency STRUCTURE	2275	1105	831
Frequency of plants containing at least one crop-specific allele	8	1	2
Total	2456	1189	924
Chi-square value for goodness of fit between STRUCTURE and crop-specific alleles	150.44 ($P < 0.001$)	80.04 ($P < 0.001$)	67.11 ($P < 0.001$)

*Crop-specific alleles used: LsA001-187, LsD103-263, LsD103-266, LsD106-191, E003-206 and E011-251.

STRUCTURE. Plants were inferred to be hybrids when they had at least one crop-specific allele. The hybrids based on crop-specific alleles were more commonly found among the STRUCTURE hybrids than in STRUCTURE nonadmixed *L. serriola* plants. However, the crop-specific allele method identified only 9% of ANGEL and CGN *L. serriola* hybrids plants, significantly fewer than STRUCTURE ($P_{\chi^2} < 0.001$). This is a very small number, even when taking into account that our necessarily strict criteria for when an allele could be considered as crop-specific, was expected to lead to conservative estimates of hybrids. Limiting the use of the crop-specific alleles to Europe, which enables using all eight alleles of Table S1 (Supporting information), as in the ANGEL data set, does not really change this situation, except for the absolute numbers (25 vs. 16 putative hybrids conforming to STRUCTURE and 16 vs. 8 not confirming to STRUCTURE, see Table 5).

Discussion

Even though crop/wild introgression is nowadays accepted as a common phenomenon, it is mostly recognized among cross-pollinating species, such as carrots (Magnussen & Hauser 2007; Rong *et al.* 2010), sunflower (Arias & Rieseberg 1994; Whitton *et al.* 1997) and chicory (Kiær *et al.* 2009). In self-pollinating species with restricted levels of cross-pollination such as lettuce, it is expected to occur (Ellstrand 2003; D'Andrea *et al.* 2008), but rarely and difficult to detect. Nevertheless, we found an occurrence of 7% of putative *L. sativa* – *L. serriola* hybrid plants from the wild habitats of *L. serriola* in Europe (recently sampled wild populations) and 9% from *L. serriola* accessions present in the CGN genebank collection. The identification of lettuce crop/wild hybrids in natural wild population implies that *L. serriola* does hybridize with *L. sativa* and that the hybrid lineages persist along with *L. serriola* nonadmixed plants. These results are different from those found in soybean (Kuroda *et al.* 2010), which is a basically self-pollinating species as well: although these authors found evidence for crop/wild hybridization, the hybrids did not persist in the natural habitats of wild soybean. Our results are consistent with previous studies in lettuce which showed that *L. sativa L. serriola* hybridization produces some hybrid lineages more vigorous and fit that the wild parent (Hooftman *et al.* 2005) and that the persistence of the hybrids depends on their relative fitness and the species outcrossing rate (Hooftman *et al.* 2007).

Lactuca sativa and *L. serriola* are so closely related that some studies have labelled them as conspecific (Koopman *et al.* 2001). Despite this close relatedness, using the ten SSR markers, STRUCTURE and InStruct differentiated the two species and identified intermediate plants, which were potential hybrids. Simko & Hu (2008) obtained similar results using STRUCTURE on a smaller set in which they could distinguish cultivated (*L. sativa*) from two wild lettuce species (*L. serriola* and *L. saligna*). The use of large data sets may improve the power and accuracy for the identification of hybrids (Burgarella *et al.* 2009), as shown here. NewHybrids recognized nearly all hybrids detected by STRUCTURE, but also several putative hybrids among the other lines. Vähä & Primmer (2006) encountered the same trend with NewHybrids, as the program classified some nonadmixed individuals as hybrids. As, in addition, NewHybrids could not handle all the available data of

our data set, NewHybrids results were used here solely for the determination of the hybrid classes. With that, one needs to be aware of possible classification errors (cf. Burgarella *et al.* 2009; Lexer *et al.* 2010). As we used a limited number of loci, the power to distinguish between various generations of selfings and backcrosses was limited (Vähä & Primmer 2006). We therefore used NewHybrids to categorize the *L. serriola* hybrid plants as identified with STRUCTURE into only two hybrid classes: all advanced selfed generations after hybridization between *L. serriola* and *L. sativa* (together represented by F7), and all advanced selfed generations after one backcross to *L. serriola* (together represented by 0-BC1S7). In a fine-scale field study of the self-pollinating species *Medicago truncatula*, Siol *et al.* (2008) found comparable results: many of the genotyped plants represented recombinant inbred lines (advanced selfed generations after a hybridization event) between the most frequently occurring highly inbred lines. In studies on the detection of spontaneous hybrids in perennial, cross-pollinating woody species, the identified hybrids often belonged to early hybrid generations such as F_2, BC_1 and BC_2 (Smulders *et al.* 2008; Schanzer & Kutlunina 2010), although the hybrid zone of *Populus alba* and *P. tremula* consisted of advanced generation hybrids (Lexer *et al.* 2010).

The higher frequency of hybrids found in the southern part of Europe for both *L. serriola* data sets could be related to the occurrence of seed multiplication in open air in the Mediterranean area (e.g. Portugal, Spain and Italy, in particular the Emilio-Romagna region), which is rare (or mostly under glass) in the north. On the other hand, the small subset of samples taken in the Netherlands near amateur gardens, one of the few places in the north where bolting and small-scale seed multiplication might occur, did not show a higher likelihood of hybrid occurrence as compared to randomly sampled populations. At such sites, bolting may only be haphazard, but it cannot be excluded that this is related to a possibly lower cross-fertilization rate under climatic conditions of northern countries.

The posterior probability approach of STRUCTURE has shown to be a good tool to identify gene flow between closely related species using molecular markers for a wide array of organisms, such as in animals, for example the carnivorous marsupial *Antechinus flavipes* (Lada *et al.* 2008), in trees, *for example* various oak species (Burgarella *et al.* 2009), and also for gene flow between crop and wild forms in, for instance, alfalfa (Greene *et al.* 2008), beet, *Beta vulgaris* (Andersen *et al.* 2005) and chicory, *Cichorium intybus* (Kiær *et al.* 2009). The fact that the software uses genotypic information encompassing all the scored alleles and their frequencies enables it to obtain a more comprehensive picture of the individuals' genetic make-up, without any previous bias of *a priori* grouping information or alleles identified as specific for any of such groups. Indeed, our trial of using the crop-specific allele approach did not work well. At best, it may lead to a conservative estimate of hybrids which logically followed from our necessarily strict definition of 'crop-specific' alleles, that is, only six alleles from five loci of a total of 315 alleles from 10 loci could at most be used as such. Moreover, about a third of the hybrids indicated by the crop-specific alleles were identified as nonadmixed plants by STRUCTURE. This could be attributed to small introgressions containing only one of the crop-specific alleles that were in the 'noise' range of the more comprehensive analysis of STRUCTURE, but it could also be due to rare coincidental occurrences of the allegedly crop-specific allele in nonadmixed wild lettuce. Indeed, recent studies using 'crop-specific' alleles often targeted more local situations with known combinations of crop cultivations and wild populations in the vicinity (*e.g.* Morrell *et al.* 2005 on introgression of sorghum into Johnson grass) or more widely different species combinations (e.g. Schulze *et al.* 2011 on garden strawberry *Fragaria* x *ananassa* and wild woodland strawberry *F. vesca* in Central Europe). Smulders *et al.* (2008), Rathmacher *et al.* (2010) and others successfully used species-specific alleles to detect gene flow and identify F1 hybrids between poplar species and hybrids. However, while crop-specific alleles are very effective in first generation hybrids, their power is lost in selfing and backcross generations, as each generation 50% of the offspring will by chance not inherit the allele and become indistinguishable from nonintrogressed plants. Thus, while useful for detecting introgression in outcrossing, long-lived perennials, crop-specific alleles are not very effective in selfing annuals in which introgression is present in advanced inbred lines in the field.

Introgression from crops to wild relatives has been connected to the invasiveness of some wild species such as Johnson grass (*Sorghum halepense*) (De Wet & Harlan 1975), *Rhododendron ponticum* (Milne & Abbott 2000) and sunflower (Rieseberg *et al.* 2007). Hooftman *et al.* (2006) suggested that introgression from *L. sativa* to *L. serriola* could be one of the reasons behind the recently observed increase of the latter in Europe, whereas D'Andrea *et al.* (2009) argued that this spread may be attributed mainly to the expansion of the favourable habitat as a result of climate warming and anthropogenic habitat disturbance and to seed dispersal because of transportation of goods. The results of this study do not support the hypothesis of Hooftman *et al.* (2006). If introgression were behind the spread of *L. serriola*, we would expect to find more putative *L. serriola* hybrids than nonadmixed *L. serriola*, particularly

in North-Western Europe where the new invasiveness of *L. serriola* was most obvious. Moreover, we would also expect to observe more hybrids among the more recently collected *L. serriola* ANGEL data set (collected between 2002 and 2005) than in the mostly older CGN genebank collection. Both expected patterns were not visible in our data. Although a number of putative *L. serriola* hybrids were found with STRUCTURE, these did not constitute the dominant proportion of the *L. serriola* plants, neither in the ANGEL nor in the CGN data sets. Moreover, hybrids were particularly rare in northern Europe. Hence, we found no evidence that crop introgression conferred an increased invasiveness to wild lettuce. Therefore, the expansion of *L. serriola* in Europe and in the Netherlands in particular resulted most likely from the combination of factors indicated by D'Andrea *et al.* (2009). Nevertheless, with lettuce being a basically self-pollinating species, the occurrence of 7% of crop/wild hybrids among natural *L. serriola* populations is relatively high and reveals a potential of transgene movement from crop to wild relatives also for self-pollinating crops. After hybridization, however, the fate of the transgene will depend on many factors including the survival and fertility of the hybrids, the fitness effect of the transgene and the relative fitness effect of the genomic region where the transgene is inserted (Stewart *et al.* 2003). The fitness effects of the genomic background in relation to environmental conditions is the topic of ongoing experimental and modelling research in a joint project of Wageningen UR, University of Amsterdam and Groningen University in the Netherlands.

Acknowledgements

Data for this publication were obtained in the context of the EU projects 'Analysis of gene flow from crop to wild forms in lettuce and chicory and its population-ecological consequences in the context of GM-crop biosafety' (ANGEL, QLK3-CT-2001-01657) and 'Molecular markers for genebanks: Application of marker technology for the improvement of ex situ germplasm conservation methodology' (PL96.2062). The data analysis was financially supported by the Netherlands' Organization for Scientific Research (NWO) through their research program 'Ecology Regarding Genetically modified Organisms' (ERGO).

References

Andersen NS, Siegismund HR, Meyer V, Jørgensen RB (2005) Low level of gene flow from cultivated beets (*Beta vulgaris* L. ssp. *vulgaris*) into Danish populations of sea beet (*Beta vulgaris* L. ssp. *maritima* (L.) Arcangeli). *Molecular Ecology*, **14**, 1391–1405.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Arias DM, Rieseberg LH (1994) Gene flow between cultivated and wild sunflowers. *Theoretical and Applied Genetics*, **89**, 655–660.

Arrigo N, Guadagnuolo R, Lappe S, Pasche S, Parisod C, Felber F (2011) Gene flow between wheat and wild relatives: empirical evidence from *Aegilops geniculata*, *Ae. neglecta* and *Ae. triuncialis*. *Evolutionary Applications*, **4**, 685–695.

Burgarella C, Lorenzo Z, Jabbour-Zahab R *et al.* (2009) Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity*, **102**, 442–452.

Chapman MA, Burke JM (2006) Letting the gene out of the bottle: the population genetics of genetically modified crops. *New Phytologist*, **170**, 429–443.

D'Andrea L, Felber F, Guadagnuolo R (2008) Hybridization rates between lettuce (*Lactuca sativa*) and its wild relative (*L. serriola*) under field conditions. *Environmental Biosafety Research*, **7**, 61–71.

D'Andrea L, Broennimann O, Kozlowski G *et al.* (2009) Climate change, anthropogenic disturbance and the northward range expansion of *Lactuca serriola* (Asteraceae). *Journal of Biogeography*, **36**, 1573–1587.

De Vries IM (1990) Crossing experiments of lettuce cultivars and species (*Lactuca* sect. *Lactuca*, Compositae). *Plant Systematics and Evolution*, **171**, 233–248.

De Vries IM (1997) Origin and domestication of *Lactuca sativa* L. *Genetic Resources and Crop Evolution*, **44**, 165–174.

De Vries IM, Van Raamsdonk LWD (1994) Numerical morphological analysis of Lettuce cultivars and species (*Lactuca* sect. Lactuca, Asteraceae). *Plant Systematics and Evolution*, **193**, 125–141.

De Wet JMJ, Harlan JR (1975) Weeds and domesticates: evolution in the man-made habitat. *Economic Botany*, **29**, 99–107.

Ellstrand NC (2003) Current knowledge of gene flow in plants: implications for transgene flow. *Philosophical Transactions of the Royal Society London, Series B*, **358**, 1163–1170.

Ellstrand NC, Schierenbeck KA (2000) Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences of the USA*, **97**, 7043–7050.

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and Introgression from domesticated plants into their wild relatives. *Annual Review of Ecoogy and Systematics*, **30**, 539–563.

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.

Falush D, Stephens M, Prtichard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Frietema de Vries FT, van der Meijden R, Brandenburg WA (1994) Botanical files on lettuce (*Lactuca sativa*): on the chance for gene flow between wild and cultivated lettuce (*Lactuca sativa* L. including *L. serriola* L., Compositae) and the generalized implications for risk-assessments on genetically modified plants. *Gorteria*, **Supplement 2**, 44 pp.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, **176**, 1635–1651.

Goudet J (1995) FSTAT (vers. 1.2): a computer program to calculate F-statistics. *Journal of Heredity*, **86**, 485–486.

Greene SL, Kisha TJ, Dzyubenko NI (2008) Conserving alfalfa wild relatives: is past introgression with Russian varieties evident today? *Crop Science*, **48**, 1853–1864.

Hooftman DAP, Oostermeijer JGB, Jacobs MMJ, Den Nijs HCM (2005) Demographic vital rates determine the performance advantage of crop–wild hybrids in lettuce. *Journal of Applied Ecology*, **42**, 1086–1095.

Hooftman DAP, Oostermeijer JGB, den Nijs JCM (2006) Invasive behaviour of Lactuca serriola (Asteraceae) in the Netherlands: spatial distribution and ecological amplitude. *Basic and Applied Ecology*, **7**, 507–519.

Hooftman DAP, Jong MJD, Oostermeijer JGB, Den Nijs HCM (2007) Modelling the long-term consequences of crop-wild relative hybridization: a case study using four generations of hybrids. *Journal of Applied Ecology*, **44**, 1035–1045.

Hooftman DAP, Hartman Y, Oostermeijer JGB, Den Nijs HCM (2009) Existence of vigorous lineages of crop-wild hybrids in Lettuce under field conditions. *Environmental Biosafety Research*, **4**, 203–217.

Kesseli RV, Ochoa O, Michelmore RW (1991) Variation at RFLP loci in *Lactuca* spp. and origin of cultivated lettuce (*L. sativa*). *Genome*, **34**, 430–436.

Kiær LP, Felber F, Flavell A *et al.* (2009) Spontaneous gene flow and population structure in wild and cultivated chicory, *Cichorium intybus* L. *Genetic Resources and Crop Evolution*, **56**, 405–419.

Koopman WJM, Jong JH, de Vries IM (1993) Chromosome banding patterns in Lettuce species (*Lactuca* sect. Lactuca, Compositae). *Plant Systematics and Evolution*, **185**, 249–257.

Koopman WJM, Zevenbergen MJ, Van Den Berg RG (2001) Species relationships in Lactuca s.l. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. *American Journal of Botany*, **88**, 1881–1887.

Kuroda Y, Kaga A, Tomooka N, Vaughan D (2010) The origin and fate of morphological intermediates between wild and cultivated soybeans in their natural habitats in Japan. *Molecular Ecology*, **19**, 2346–2360.

Lada H, Mac Nally R, Taylor AC (2008) Distinguishing past from present gene flow along and across a river: the case of the carnivorous marsupial (*Antechinus flavipes*) on southern Australian floodplains. *Conservation Genetics*, **9**, 569–580.

Langevin SA, Clay K, Grace JB (1990) The incidence and effects of hybridization between cultivated rice and its related weed red rice (*Oryza sativa* L.). *Evolution*, **44**, 1000–1008.

Lebeda A, Doležalová I, Křístková E, Mieslerová B (2001) Biodiversity and ecogeography of wild *Lactuca* spp. in some European countries. *Genetic Resources and Crop Evolution*, **48**, 153–164.

Lebeda A, Doležalová I, Astley D (2004a) Representation of wild *Lactuca* spp. (Asteraceae, Lactuceae) in world genebank collections. *Genetic Resources & Crop Evolution*, **51**, 167–174.

Lebeda A, Doležalová I, Feráková V, Astley D (2004b) Geographical distribution of wild *Lactuca* species (Asteraceae, Lactuceae). *Botanical Review*, **70**, 328–356.

Lexer C, Joseph JA, van Loo M *et al.* (2010) Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. *Genetics*, **186**, 699–712.

Magnussen LS, Hauser TP (2007) Hybrids between cultivated and wild carrots in natural populations in Denmark. *Heredity*, **99**, 185–192.

Milne RI, Abbott RJ (2000) Origin and evolution of invasive naturalized material of *Rhododendron ponticum* L. in the British Isles. *Molecular Ecology*, **9**, 541–556.

Morrell PL, Williams-Coplin TD, Lattu AL, Bowers JE, Chandler JM, Paterson AH (2005) Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. *Molecular Ecology*, **14**, 2143–2154.

Pilson D, Prendeville HR (2004) Ecological effects of transgenic crops and the escape of transgenes into wild populations. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 149–174.

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

Rathmacher G, Niggemann M, Köhnen M *et al.* (2010) Short-distance gene flow in *Populus nigra* L. accounts for small-scale spatial genetic structures: implications for in situ conservation measures. *Conservation Genetics*, **11**, 1327–1338.

Rieseberg LH, Kim SC, Randell RA *et al.* (2007) Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, **129**, 149–165.

Rong J, Janson S, Umehara M, Ono M, Vrieling K (2010) Historical and contemporary gene dispersal in wild carrot (*Daucus carota* ssp. *carota*) populations. *Annals of Botany*, **106**, 285–296.

Saisho D, Purugganan MD (2007) Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World. *Genetics*, **177**, 1765–1776.

Schaminée JHJ, van Duuren L, de Bakker AJ (1992) Europese en mondiale verspreiding van Nederlandse vaatplanten. *Gorteria*, **18**, 57–96.

Schanzer IA, Kutlunina NA (2010) Interspecific hybridization in wild roses (*Rosa* L. sect. Caninae DC.). *Biology Bulletin*, **37**, 480–488.

Schulze J, Stoll P, Widmer A, Erhardt A (2011) Searching for gene flow from cultivated to wild strawberries in Central Europe. *Annals of Botany*, **107**, 699–707.

Scurrah M, Celis-Gamboa C, Chumbiauca S *et al.* (2008) Hybridization between wild and cultivated potato species in the Peruvian Andes and biosafety implications for deployment of GM potatoes. *Euphytica*, **164**, 881–892.

Simko I, Hu J (2008) Population structure in cultivated lettuce and its impact on association mapping. *Journal of the American Society for Horticultural Science*, **133**, 61–68.

Siol M, Prosperi J, Bonnin I, Ronfort J (2008) How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. *Heredity*, **100**, 517–525.

Smulders MJM, Beringen R, Volosyanchuk R *et al.* (2008) Natural hybridization between *Populus nigra* L. and P. x canadensis Moench. Hybrid offspring competes for niches along the Rhine river in the Netherlands. *Tree Genetics and Genomes*, **4**, 663–675.

Snow AA, Andow DA, Gepts P *et al.* (2005) Genetically engineered organisms and the environment: current status and recommendations. *Ecological Applications*, **15**, 377–404.

Stewart Jr CN, Halfhill MD, Warwick SI (2003) Transgene introgression from genetically modified crops to their wild relatives. *Nature Reviews Genetics*, **4**, 806–817.

Storme V, Vanden Broeck A, Ivens B *et al.* (2004) Ex-situ conservation of Black poplar in Europe: genetic diversity in

nine gene bank collections and their value for nature development. *Theoretical and Applied Genetics*, **108**, 969–981.
Thompson RC, Whitaker TW, Bohn GW, Van Horn CW (1958) Natural cross-pollination in lettuce. *American Society for Horticultural Science*, **72**, 403–409.
Truco MJ, Antonise R, Lavelle D *et al.* (2007) A high-density, integrated genetic linkage map of lettuce (*Lactuca* spp.). *Theoretical and Applied Genetics*, **115**, 735–746.
Ureta MS, Cantamutto M, Carrera A *et al.* (2008) Natural hybrids between cultivated and wild sunflowers (*Helianthus* spp.) in Argentina. *Genetic Resources and Crop Evolution*, **55**, 1267–1277.
Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.
Van de Wiel CCM, Arens P, Vosman B (1999) Microsatellite retrieval in lettuce (*Lactuca sativa* L.). *Genome*, **42**, 139–149.
Van de Wiel CCM, Van der Linden G, Den Nijs H *et al.* (2003) An EU project on gene flow analysis between crop and wild forms of lettuce and chicory in the context of GMO biosafety: first results in lettuce. In: *EUCARPIA leafy vegetables 2003. Proceedings of the EUCARPIA Meeting on Leafy Vegetables Genetics and Breeding, Noordwijkerhout, Netherlands, 19–21 March 2003* (eds Van Hintum TJL, Lebeda A, Pink DA and Schut JW), pp. 111–116. Centre for Genetic Resources, the Netherlands (CGN), Wageningen.
Van de Wiel CCM, Groot M, Den Nijs H (2005) Gene flow from crops to wild plants and its population-ecological consequences in the context of GM-crop biosafety, including some recent experiences from lettuce. In: *Environmental Costs and Benefits of Transgenic Crops* (ed. Wesseler J), pp. 97–110. Springer, Dordrecht.
Van de Wiel CCM, Sretenović Rajičić T, Van Treuren R *et al.* (2010) Distribution of genetic diversity in wild European populations of prickly lettuce (*Lactuca serriola*): implications for plant genetic resources management. *Plant Genetic Resources-Characterization and Utilization*, **8**, 171–181.
Van Hintum TJL (2003) Molecular characterisation of a lettuce germplasm collection. In: *Eucarpia Leafy Vegetables 2003, Proceedings of the Eucarpia Meeting on Leafy Vegetables, Noordwijkerhout, The Netherlands, 19–21 March 2003* (eds Van Hintum TJL, Lebeda A, Pink D and Schut JW), pp. 99–104. Centre for Genetic Resources, The Netherlands (CGN), Wageningen, The Netherlands.
Van Treuren R, Van Hintum TJL, Van de Wiel CCM (2008) Marker-assisted optimization of an expert-based strategy for the acquisition of modern lettuce varieties to improve a genebank collection. *Genetic Resources and Crop Evolution*, **55**, 319–330.
Warwick SI, Légère A, Simard M-J, James T (2008) Do escaped transgenes persist in nature? The case of an herbicide resistance transgene in a weedy Brassica rapa population. *Molecular Ecology*, **17**, 1387–1395.
Warwick SI, Beckie HJ, Hall LM (2009) Gene flow, invasiveness, and ecological impact of genetically modified crops. *Annals of the New York Academy of Sciences*, **1168**, 72–99.
Westman AL, Levy BM, Miller MB *et al.* (2001) The potential for gene flow from transgenic crops to related wild species: a case study in strawberry. *Acta Horticulturae*, **560**, 527–530.
Whitton J, Wolf DE, Arias DM *et al.* (1997) The persistence of cultivar alleles in wild populations of sunflowers five generations after hybridization. *Theoretical and Applied Genetics*, **95**, 33–40.

This study is part of the PhD thesis of BU, supervised by CCMvdW, MJMS and RGFV. BU is interested in plant breeding using state-of-the-art-tools from molecular biology and statistics, such as molecular markers and QTL mapping. CCMvdW, MJMS and RGFV are researcher, group leader and head, respectively, of Wageningen UR Plant Breeding. CCMvdW is interested in genetic diversity and gene flow in relation to environmental safety and sustainability of GM crops. MJMS is interested in the use of various methods to study neutral and functional (agro)biodiversity. RGFV is interested in exploiting genetic variation to improve crop plants. Part of the molecular data was gathered as part of the PhD thesis of LD'A, supervised by FF. LD'A is interested in environmental issues linked to GM crop cultivation and is expanding his interests in the agricultural– ecological field of research as well as in fields such as ecological economics and agroecology. FF is presently director of the BotanicalMuseum and Gardens of Canton Vaud in Lausanne and Pont de Nant and is interested in gene flow between crops and wild relatives. DAPH and HCMdN were involved in collecting plant material. DAPH is researcher at CEH and is interested in empirical testing and statistical modelling of the consequences of gene flow for plant populations, ranging from genetics to landscape dynamics. HCMdN is guest-researcher at IBED and member of COGEM, the Netherlands Commission on Genetic Modification, and is interested in experimental plant systematics and conservation biology of plants.

Data accessibility

Detailed information on the used data (origin of the samples and genotype of each individual based on the 10 microsatellites) may be found in the online supplementary material I.

Supporting information

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

Fig. S1 STRUCTURE *Q* graph of all datasets at (A) $K = 3$ and (B) $K = 4$: At higher K (>2) *L. serriola* remains distinct from *L. sativa*.

Table S1 Frequency of ''crop specific'' alleles in lettuce datasets.

Appendix S1 Supplementary material I.

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.