MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 4182–4194 doi: 10.1111/j.1365-294X.2011.05258.x

Recent long-distance transgene flow into wild populations conforms to historical patterns of gene flow in cotton (*Gossypium hirsutum*) at its centre of origin

A. WEGIER,*† A. PIÑEYRO-NELSON,*‡ J. ALARCÓN,§ A. GÁLVEZ-MARISCAL, ¶ E. R. ÁLVAREZ-BUYLLA*‡ and D. PIÑERO*

**Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado postal 70-725, CP 04510, México DF, México, †Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Progreso 5, Coyoacán, 04010, México DF, México, ‡Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Apartado postal 70-725, CP 04510, México DF, México, §Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Liga Periférico-Insurgentes Sur 4903, Parques del Pedregal, Tlalpan 14010, México DF, México, ¶Departamento de Alimentos y Biotecnología, Facultad de Química, Universidad Nacional Autónoma de México, CP 04510, México DF, México*

Abstract

Over 95% of the currently cultivated cotton was domesticated from *Gossypium hirsutum*, which originated and diversified in Mexico. Demographic and genetic studies of this species at its centre of origin and diversification are lacking, although they are critical for cotton conservation and breeding. We investigated the actual and potential distribution of wild cotton populations, as well as the contribution of historical and recent gene flow in shaping cotton genetic diversity and structure. We evaluated historical gene flow using chloroplast microsatellites and recent gene flow through the assessment of transgene presence in wild cotton populations, exploiting the fact that genetically modified cotton has been planted in the North of Mexico since 1996. Assessment of geographic structure through Bayesian spatial analysis, BAPS and Genetic Algorithm for Rule-set Production (GARP), suggests that *G. hirsutum* seems to conform to a metapopulation scheme, with eight distinct metapopulations. Despite evidence for long-distance gene flow, genetic variation among the metapopulations of *G. hirsutum* is high (He = 0.894 ± 0.01). We identified 46 different haplotypes, 78% of which are unique to a particular metapopulation, in contrast to a single haplotype detected in cotton cultivars. Recent gene flow was also detected (m = 66/270 = 0.24), with four out of eight metapopulations having transgenes. We discuss the implications of the data presented here with respect to the conservation and future breeding of cotton populations and genetic diversity at its centre of crop origin.

Keywords: *Gossypium hirsutum*, long distance gene flow, metapopulations, Mexico, transgene flow

Received 1 October 2010; revision received 6 July 2011; accepted 15 July 2011

Introduction

The complexes of wild and cultivated varieties of crop plants at their centres of crop origin and/or diversity (hereafter, CCO) provide useful systems for addressing fundamental questions on population structure, genetics, and specifically, gene flow dynamics (e.g. maize to teosinte; Baltazar *et al.* 2005; Ellstrand *et al.* 2007; the beet family; Bartsch *et al.* 1999; Viard *et al.* 2004; Fénart *et al.* 2007; Arnaud *et al.* 2009; or *Brassica* spp. Jørgensen & Andersen 1994; Snow *et al.* 1999). In cases where genetically modified varieties have been released at the CCO, transgenes become useful markers for addressing ongoing patterns, dynamics, and pervasiveness of gene flow (maize, van Heerwaarden *et al.* 2009; *Cucurbita*, Sasu *et al.* 2009; *Sorghum*, Sahoo *et al.* 2010). At the same time, these cases become particularly relevant for assessing the

Correspondence: Ana Wegier, Fax: (+52 55) 36 26 87 00 Ext. 104
E-mail: awegier@gmail.com
Present address: CENID-COMEF, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Progreso 5, Coyoacán, 04010, México DF, México.

general viability of GMO cultivation when there is potential for transgene flow into wild relatives at the CCO.

In spite of the effects of recent gene flow involving transgenes or other genetic elements, historical gene flow may still have a dramatic impact on population genetic structure (Ehrlich & Raven 1969). It may counteract the effects on effective population size of drift and inbreeding (Ebert *et al.* 2002), but may also constrain population differentiation by homogenizing the gene pool (Slatkin 1987). Gene flow estimation has historically relied on estimates of *Nm* (number of migrants per generation) and F_{st} (Fixation index; a measure of population differentiation). However, both of these parameters have been developed based on simplified and typically unrealistic population models (Whitlock & McCauley 1999; Paetkau *et al.* 2004) that assume, for example, that populations are at equilibrium (Broquet & Petit 2009).

In contrast, the use of haplotype networks and genetic covariance estimates, such as those used in Popgraph analyses, can provide information regarding the historical and spatial relationships among genotypes (Dyer 2009). For instance, historical gene flow patterns can be inferred from haplotype networks that connect each particular haplotype through mutational steps. This enables assignment of extant haplotypes to an ancestral population, while differentiating between ancestral polymorphisms and migration. This distinction is particularly useful when analysing species that have diversified or diverged quite recently, as is the case for the majority of cultivars (Londo *et al.* 2006). On the other hand, Popgraph draws from tools generated by landscape genetics that allow for the differentiation between isolation by distance and long distance migrations, which are phenomena that can underlie genetic differentiation among populations (Dyer & Nason 2004). These approaches explicitly incorporate geographical information to assess the contribution of physical space in structuring genetic diversity (Manel *et al.* 2003; Dyer 2009).

In the present study, we complement these types of historical gene flow analyses with estimates of ongoing gene flow using transgenes. While gene flow estimation is instrumental in the analysis of the genetic structure of populations, it should be complemented with a direct assessment of pollen and seed dispersal rates that impact on the natural patterns of gene flow. Otherwise, the consequences of dispersal-related life history variation among populations—and, hence, gene flow itself—-will remain poorly quantified (e.g. Palstra *et al.* 2007). Therefore, we have also pursued the analysis of landscape features that can impact the genetic structure of populations by documenting the metapopulation structure of cotton in Mexico.

Metapopulations are assemblages of populations that exist in a balance between extinction and colonization (Levins 1969; Hanski & Gaggiotti 2004 and references therein). For plants, several criteria have been proposed that further constrain this metapopulation concept (Hanski 1998; Freckleton & Watkinson 2002), including: (i) that suitable metapopulation habitats are in spatially separated patches; (ii) that all patches can become extinct but they cannot do so at the same time; and (iii) that recolonization of each patch after local extinction is possible (Honnay *et al.* 2009).

The complex of wild and cultivated cotton populations in Mexico is an ideal system with which to address the role of metapopulation dynamics on recent and historical gene flow patterns, and on the genetic structure of populations. These studies are also instrumental for breeding and conservation programs for crops at their CCO. The germplasm of current cultivated cotton originated in Mesoamerica, where it was semi-domesticated in pre-Hispanic times (Tehuacán Valley, Mexico, dated around 5500–4300 BP; Smith & Stephens 1971). Previous studies used allozymes and RFLP data to identify possible venues of cotton domestication and to assess broad range genetic diversity (Wendel & Albert 1992; Brubaker & Wendel 1994). However, although cultivated cotton varieties are the most important source of natural fibre and the third source of oil in the world (FAOSTAT 2009), only two varieties (*G. hirsutum* var. *yucatanense*; called TX2094 and Deltapine 14; Delta and Pine Land Co; Applequist *et al.* 2001) have been used as reference for wild germplasm. Thus, broadening the genetic studies of wild populations of *G. hirsutum* will increase the success of breeding strategies focused on generating varieties adapted to extreme environments.

The *Gossypium* genus originated from African relatives between 12.5 (Seelanan *et al.* 1997) and 25 (Wendel & Albert 1992; Wendel *et al.* 2010) million years ago, and its salt-tolerant seeds enabled its spread around the world (Stephens 1966; Seelanan *et al.* 1997). Only four out of more than fifty *Gossypium* species have been domesticated (Wendel *et al.* 2009): two diploids in Asia and Africa (*G. herbaceum* and *G. arboreum*) and two tetraploids in America (*G. hirsutum* and *G. barbadense*). Current diploid and allopolyploid *Gossypium* species on the American continent cannot hybridize amongst themselves (Beasley 1940, 1942). Cotton is mainly self-pollinated, although cross-pollination may rarely occur (Stephens & Finkner 1953; Simpson 1954; McGregor 1976), and gene flow occurs via seed dispersal by water (Stephens 1966), and probably by wind and birds. In Mexico, GM cotton has been cultivated since 1996 and 172 000 ha were approved for sowing in 2009 (SAGARPA 2010). Despite the extent of GM cotton cultivation, the dispersal of transgenes into non-GM and wild cotton has not yet been evaluated.

Given the complex history of the *Gossypium* genus and its capability for long distance migration, we first evaluated the geographical structuring of *G. hirsutum* populations in Mexico by generating a potential distribution based on climatic data. We hypothesized that geographic barriers have affected long distance gene flow among wild *G. hirsutum* populations, rendering a genetic structure that does not conform to an isolation-by-distance pattern across the area as a whole. We then documented historical gene flow using chloroplast microsatellite data to construct a haplotype network. Lastly, we used transgenes as markers to assess whether recent gene flow has taken place and if its patterns and dynamics conform to our historical inferences.

Materials and methods

Assessment of wild cotton populations and modelling of a potential distribution map

We selected populations of wild *Gossypium hirsutum* to be collected for this work by first performing an analysis of hundreds of historical specimens at the *MEXU* National Herbarium and *XAL* Herbarium. Twenty accessions were used that were clearly referenced as wild specimens and whose geographical reference fell within the formerly established natural habitats of this species. Concomitantly, we used the collections made by Paul A. Fryxell between 1968 and 1975 to guide our field search for wild populations. The specimens collected by Fryxell had clear features of wild cotton, as well as a precise description of both the habitat and location. Based on previous reports (Fryxell 1979; Wendel & Albert 1992; Applequist *et al.* 2001), we used the following objective criteria to classify a cotton plant as wild: (i) it is present in the expected habitat and distribution for the species' wild populations; (ii) it is a perennial shrub or tree, and (iii) its fruits have less than 22% lint content. We also delimited our unit of study, considering a population as comprised by a set of individuals that may potentially cross-pollinate among themselves and that are set at a distance of a maximum of 14 km among them. This distance criterion was set as a conservative limit, because this is the maximum pollinator (honeybee) movement range reported to date (Beekman & Ratnieks 2000).

We characterized the ecological niche for this species, based on 185 collection points of wild cotton plants surveyed between 2002 and 2007. We used the niche model proposed by Wiley (Wiley *et al.* 2003) and analysed our data using the GARP program (Genetic Algorithm for Rule-set Production; Scachetti-Pereira 2001), which incorporated 23 bioclimatic covers from Worldclim, with a convergence limit of 0.01%, a 5% of omission, and a 10% commission threshold. Models were selected using the methodology proposed by Anderson *et al.* (2003).

The potential distribution map of *G. hirsutum* wild populations in Mexico was delimited through comparison with cartography from the Biogeographic Regions of CONABIO (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad 1997). The predicted areas of distribution of wild cotton were validated through field inspections of places diagnosed to contain *G. hirsutum* wild populations according to the potential distribution maps, but where no former collections had been undertaken or where no entries were available at any database consulted. With these data, we analysed the population structure using a metapopulation scheme (see 'Results' section).

Cotton seed collection

Cotton seeds were collected between 2002 and 2008 in the identified wild populations of this species. The size of surveyed populations varied in number from 4 to 24 plants, with 1 up to <14 km separating individuals within a population. In total, 336 individual plants distributed in 36 populations were collected (Fig. 1).

Additionally, seeds from commercial cotton cultivars from Sonora, Mexicali, Chihuahua (Mexico), Texas, Virginia (USA), Argentina, Brazil, India and Egypt, were used to assess genetic diversity. We also collected seed from populations present outside the potential distribution area: Cuautla (18.89 N, −99.97 W), Tepoztlán (18.97 N, −99.09 W), Cuernavaca (18.87 N, −99.205 W; in the state of Morelos), Durango (23.18 N, −104.52 W) and Sonora (28.80 N, −110.57 W). All of these were considered feral populations because they have more than 25% of lint and are far away from potential distribution areas of wild cotton.

Laboratory procedures

DNA and chloroplast microsatellite analyses. Collected cotton seeds were germinated in growth chambers with a 12 h/30 °C light and 12 h/20 °C darkness regime, in 80–90% humidity. Genomic DNA was isolated from young seedling leaves using a modified CTAB method from Sul & Korban (1996; see Table S1).

DNA sequences were amplified through PCR using two specific primer sets for *G. hirsutum* chloroplast microsatellites (AF351292 $(GAA)_9$ and AF351313 $(CA)_{12}$, from Reddy *et al.* 2001). Additionally, ten PCR primer sets were used for the analysis of simple sequence

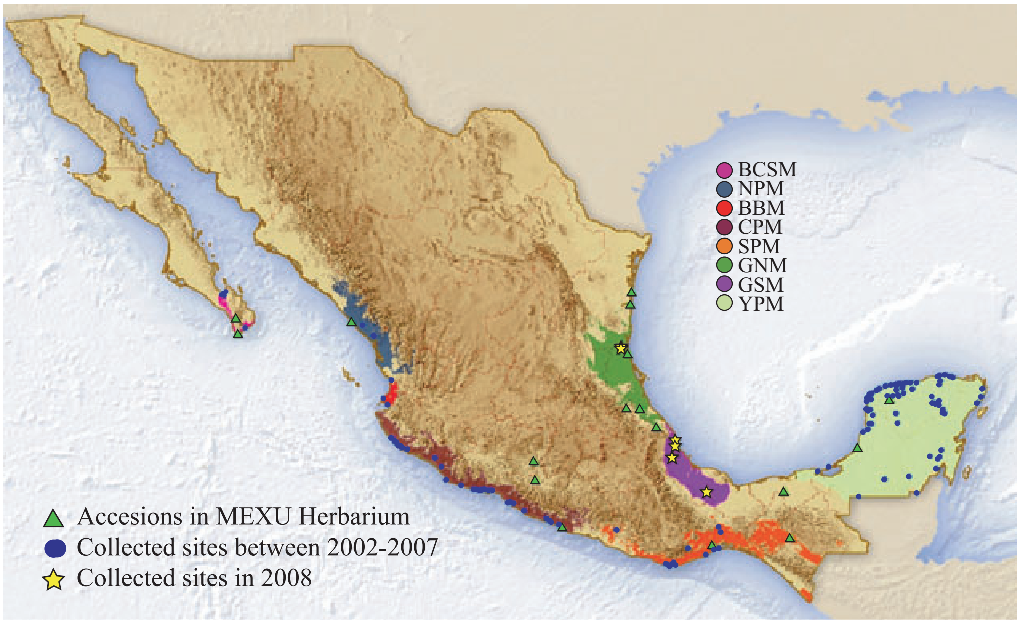
BCSM
NPM
BBM
CPM
SPM
GNM
GSM
YPM
Accesions in MEXU Herbarium
Collected sites between 2002-2007
Collected sites in 2008

Fig. 1 Map showing collection sites, potential distribution area and metapopulations of *G. hirsutum* in Mexico. Symbols: green triangles: cotton collections identified by Fryxell at the MEXU herbarium; blue circles: 2002–2007 cotton collections; yellow stars: 2008 collections discovered with the use of the potential distribution map. Metapopulations are coded as follows: Baja California Sur (BCSM): fuchsia; North Pacific (NPM): grey; Banderas Bay (BBM): red; Central Pacific (CPM): burgundy; South Pacific (SPM): orange; Gulf North (GNM): dark green; Gulf South (GSM): purple; and Yucatán Peninsula (YPM): lime.

repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms (CCMP1–CCMP10 from Weising & Gardner 1999). The PCR procedure and individual conditions are shown in Table S1. DNA fragments were sequenced on an ABI Prism 3730xl Analyzer at the High-Throughput Sequencing and Genotyping Unit in the University of Illinois.

Immunoassays to detect the presence of transgenes in wild cotton populations. A total of 270 individual cotton seeds from 36 populations ($N \geq 20$ seeds per population; see Table 1) were individually analysed for transgene presence via immunoassays for the most common recombinant proteins present in cultivated cotton for which ELISA kits were available (Cry1Ab/Ac, Cry2A, CP4-EP-

Table 1 Presence of recombinant proteins in *G. hirsutum* metapopulations

Metapopulation	Populations	*N*	Positive seeds	Positive 1 protein	Positive 1 + proteins
BCS (S of BCS)	2	17	0	0	0
North Pacific (Center and S of Sinaloa and N of Nayarit)	3	37	25	19	6
Banderas Bay (SW Nayarit and NW of Jalisco)	2	15	0	0	0
Center Pacific (Coastal line of C and S of Jalisco, Colima, Michoacán and NW and C of Guerrero)	6	24	0	0	0
South Pacific (SE of Guerrero, Coastal line of Oaxaca, CW, C and South tip of Chiapas)	8	44	13	13	0
Yucatán Peninsula (Quintana Roo, Yucatán, Campeche and NE and E of Tabasco)	11	88	0	0	0
Gulf South (C and SE of Veracruz)	3	21	14	12	2
Gulf North (N of Veracruz, E of San Luis Potosí and S of Tamaulipas)	1	24	14	0	14
Total	36	270	66	44	22

The region comprised within each metapopulation is described in parentheses; the number of wild cotton populations in each metapopulation is presented in column two. Symbols: *N*: total number of seeds analysed per metapopulation; positive: total number of seeds positive for recombinant protein presence; positive 1 protein: number of seeds positive for only one recombinant protein; positive 1 + protein: number of seeds positive for more than 1 and up to 4 different recombinant proteins (see text for a complete description).

SPS and PAT/Bar). The embryo of each seed was separated from its seed coat and divided into four pieces with a surgical knife. Each piece was placed in a 2-mL microcentrifuge tube for separate homogenization with an appropriate volume of extraction buffer. Each sample was analysed using duplicate assays in each ELISA plate. Immunoassays were conducted according to the manufacturer's instructions. The ELISA plates were read in a spectrophotometer at 450 nm for proteins PAT/bar, Cry2A and CP4-EPSPS-event NK603 (Envirologix™ plates) and at 650 nm for proteins CP4-EPSPS and Cry1Ab/Ac (Agdia® plates).

We considered a sample to be positive only when its absorbance was equal to or above a reading three standard deviations above the average intensity of all negative controls and blank samples. At least one duplicate of a blank (extraction buffer), one negative control, and one positive control were included in each ELISA plate.

Data analyses

Molecular diversity. We determined the number and frequency of all unique chloroplast DNA haplotypes and estimated molecular diversity using Arlequin v3.5 (Excoffier & Lischer 2010). We used the rarefaction approach (using ADZE; Szpiech *et al.* 2008) to see if heterogeneous population sizes could affect the estimation of genetic diversity among populations and also to generate estimates that would be comparable among different populations (Petit *et al.* 1998; Kalinowsky 2004).

Genetic structure and gene flow analyses. We examined population structure by performing a Bayesian spatial analysis using the program BAPS 5.1 (Corander *et al.* 2008), which uses stochastic optimization to find the optimal partition. Simulations were run from $K = 2$ to $K = 10$ with 100 replicates for each K.

We sought evidence for isolation by distance and/or long-distance dispersal events using Population Graph (GeneticStudio software; Dyer 2009). This is a graph-theoretic approach that analyses how genetic variation is distributed across the investigated landscape, by plotting migration and enabling the assessment of the dependence or independence of evolutionary trajectories among populations. Within a graph, populations are represented as nodes and the genetic covariation among populations determines the topology. The pattern of connections between populations is estimated conditional on the entire data set. The pattern can be used to test for isolation-by-graph-distance, where in an extreme case, if covariance between two populations equals zero, no connection is drawn (IBGD; Dyer & Nason 2004). Plotting the Population Graph onto a map also allows the inferred population pairs to have 'extended edges', 'normal edges', or 'compressed edges', which imply that genetic distance is either higher, equal to, or lower, respectively, than the one expected by geographic data (Dyer 2009).

We investigated the evolutionary history and relationships among the haplotypes found in this study and differentiation of the ancestral polymorphism and gene flow by constructing a minimum-spanning network of haplotypes using TCS 1.21 (Clement *et al.* 2000). We used the methods described by Templeton & Sing (1993) to break loops (ambiguous connections) within our network, while using predictions derived from coalescence theory (reviewed in Rosenberg & Nordborg 2002).

Distances between GM cotton release sites and wild cotton populations. Mexico was divided into over 80 000 hexagons, 25 km^2 each, to compare areas against experimental release centres. Centroids of these hexagons were used to calculate the distance between the release sites and the potential distribution model, with an error of 25 km. The sites where permits to release genetically modified cotton in Mexico have been granted (from 1996 to 2008) were plotted on a map of Mexico, under the assumption that all plots approved were actually planted (Fig. 4a). The minimum distance separating a granted GM cotton release site from all populations of wild cotton was determined (Table 2).

Results

Wild populations of G. hirsutum in Mexico: potential distribution and actual metapopulation structure

A potential distribution map was generated using computational and geographic tools (GARP). This map was based on a comprehensive survey of existing wild *G. hirsutum* populations comprising 185 collection points (recorded between 2002 and 2007) distributed in 36 populations (blue points in Fig. 1). The potential distribution is plotted in Fig. 1 and represents those areas that had over 75% of confidence of translating into the actual wild cotton distribution, according to our survey data. Thus, the actual distribution area for this species may possibly be even broader than that considered here. Nevertheless, the fact that all predicted populations were either corroborated or led to the finding of new populations helped us to validate the precision of the ecological niche prediction model used here. The potential distribution map identified seven new populations along the Gulf of Mexico in 2008 (yellow stars in Fig. 1). In previous years, without the guidance of this model, efforts to find wild populations in this area had proved unsuccessful.

During the 7-year fieldwork period (2002–2008), we observed that 85% of wild cotton populations were in coastal ecosystems and some were in low dry forests. Cotton plants were in populations of 4 to 20 individuals (5 was the mode).

The spatial distribution and the ecological setting of the populations investigated here suggest the existence of eight discrete bioclimatic areas. These are separated by intermediate zones that lack adequate climatic and ecological conditions for *G. hirsutum* to grow, and that effectively form geographical barriers to seed and pollen flow. Each discrete area described here is considered to be a metapopulation because cotton plant populations were discontinuous due to the discrete occurrence of favourable habitats. Furthermore, each metapopulation was separated from one another by at least 150 km or was isolated by evident geographical barriers.

The eight metapopulations proposed here are: Baja California Sur (BCSM), North Pacific (NPM), Banderas Bay (BBM), Center Pacific (CPM), South Pacific (SPM), Yucatán Peninsula (YPM), Gulf South (GSM) and Gulf North (GNM; Fig. 1). Although we lack quantitative dynamic data for all of the populations surveyed, the number of plants per population, as well as the number of populations that form a metapopulation, varied substantially (Table 1). In the northern part of the country, the maximum number of populations per metapopulation is three (BCSM, NPM, BBM and GSM). In the south of Mexico, three metapopulations have six, eight, and eleven populations (CPM, SPM and YPM, respectively). With regard to suitable habitats for cotton growth within metapopulations, the YPM has the largest contiguous range of suitable habitats and bears the largest populations. It is also the most genetically diverse.

Genetic variation, historical gene flow, and population structure of G. hirsutum *in Mexico*

Overall, genetic variation among wild metapopulations of *G. hirsutum* is high (He = 0.894 ± 0.01). We found a total of 46 haplotypes, 78% of which are unique to a particular metapopulation (Fig. 2). The highest haplotype diversity was found in BBM (0.94) and YPM (0.93; for haplotype diversity between metapopulations, see Table S2). The remaining metapopulation diversity ranges between values of 0.6 and 0.8, except for GSM, which is exceptionally low (0.34). In contrast, the analysed commercial cotton seeds and inferred feral populations have only one haplotype (number 2; Fig. 2a). The only exception to this trend is the feral population in Cuernavaca, Morelos, where two haplotypes were found (2 and 25; Fig. 2a).

The 46 haplotypes found in this study group into six distinct lineages, of which those of BBM and YPM are well differentiated (Fig. 2a). This haplotype network has a complex topology, where some populations with unique lineages have haplotypes that are not sampled,

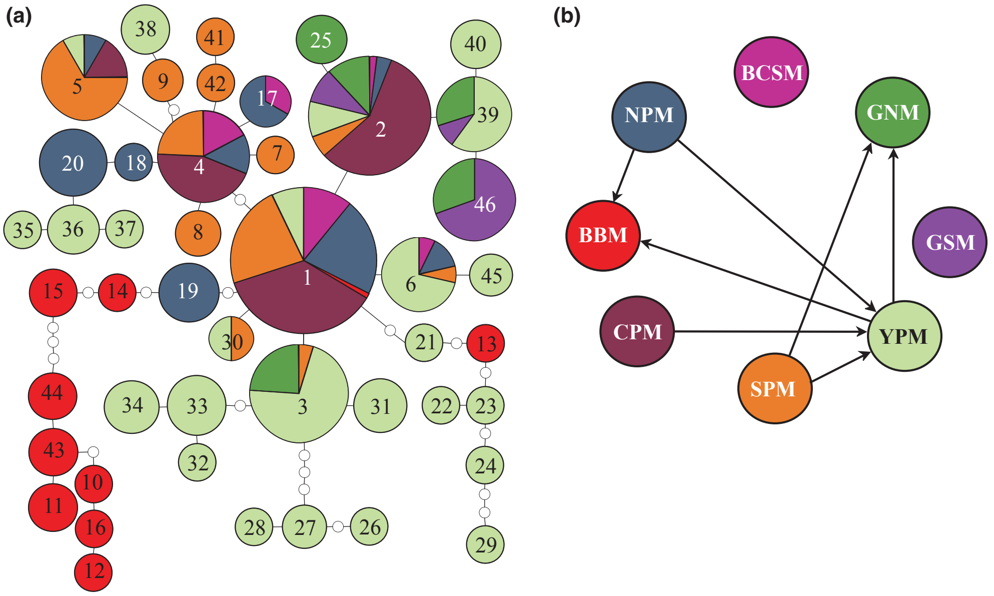

Fig. 2 Haplotype network and historical gene flow in wild *G. hirsutum* metapopulations. (a) Haplotype network. Haplotypes documented in this work are depicted in circles; sizes of nodes show the frequency of a particular haplotype while colours represent the presence of a particular haplotype within each metapopulation. (b) Historical gene flow patterns among metapopulations, as inferred from the haplotype network (metapopulation colour-codes and labels are as in Fig. 1).

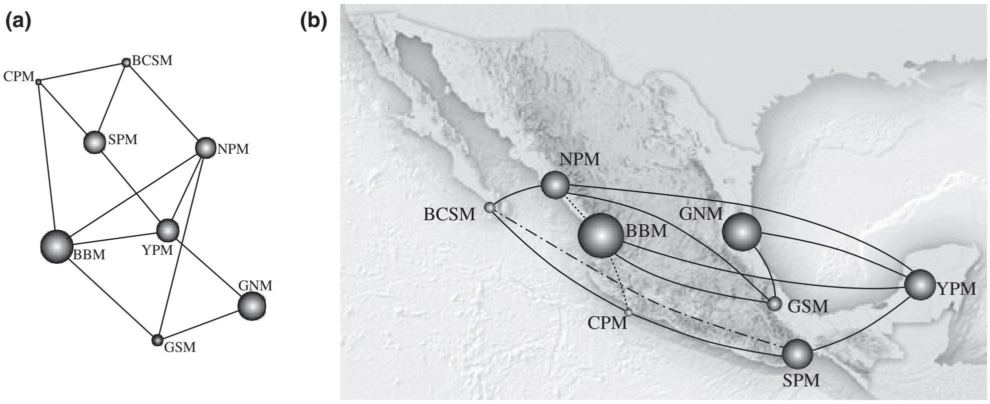

Fig. 3 *Popgraph* analysis showing significant connections among wild cotton metapopulations. (a) Three-dimensional *Popgraph* representing the genetic covariance among metapopulations of *G. hirsutum,* based on chloroplast markers. The length of a line between any two metapopulations is proportional to their covariance; within-metapopulation genetic variance is proportional to sphere size. (b) Geographic distances among metapopulations and their relation with the *Popgraph* analysis shown in (a). Edges that are significantly longer (– - - - –), shorter (—) or congruent (–––) with the predicted genetic covariance with respect to geographic distance are plotted. Names are as in Fig. 1.

while shared haplotypes among several populations appear to be ancestral. Derived haplotypes are generally unique. The unique haplotype of cultivated cotton (2) is shared and frequent in almost all metapopulations. Haplotypes 35, 36, and 37 show ancient gene flow, while 38 and 5 seem to have recently migrated from the YPM; lastly, haplotype 13 shows evidence of ancestral gene flow (Fig. 2b).

We estimated the genetic structure among the surveyed wild cotton populations by modelling our data using the BAPS approach. We found that the optimal number of clustered groups was eight; thus, the description of each cluster by this algorithm is consistent with the metapopulation scheme derived from the potential distribution analysis used above. Furthermore, the rarefaction approach was consistent with the Population Graph tool, as the diameter representing the genetic variation is not correlated with the sample size per population (Fig. 3).

Given the inferences of genetic variation and metapopulation genetic diversity presented above, which are the result of historical events, we addressed the question of whether contemporary long distance gene flow has taken place, by evaluating transgene flow.

Recent long distance gene flow: presence of transgenes in wild cotton metapopulations

The potential for long and short-distance ongoing transgene flow that could be occurring from GM cotton plants to native wild cotton populations was evaluated through plotting the frequency distribution of the distance between the GM cotton parcels and the nearest wild cotton population (Table 2). In this analysis, we found that 1.4% of 5985 permits to sow GM cotton issued by the pertinent Mexican authority between 1996 and the beginning of 2008, fall within the area of distribution of two metapopulations of wild cotton (NPM and GNM), while 4.2% are within a 300-km radius from three metapopulations (NPM, GNM and GSM). The remaining 94.4% of GM field releases approved are over 300 km apart from all wild cotton metapopulations (Table 2).

We identified actual transgene flow by assessing the presence of recombinant proteins in wild cotton populations through ELISA tests. The immunoassays yielded 66 positive seeds out of 270 seeds tested (24.4%) for at least one recombinant protein (Table 1). These positive cases were distributed among four metapopulations (Fig. 4): NPM (25/37; 67.6%), GNM (14/24; 58.3%), GSM (14/21; 66.7%) and surprisingly, SPM (13/44; 29.5%). The latter is at a lineal distance of 755 km from the southernmost and nearest approved GM cotton plot. Furthermore, 3 out of 3 populations comprising the NPM had positive testing plants for transgene presence; 1 out of 1 in GNM; 2 out of 3 in GSM and 3 out of 7 in the SPM. Interestingly, two-thirds of all positive samples yielded positive results for a single recombinant protein, while one-third did so for two and up to four different transgene-codified proteins (Table 1).

Of the 66 positive seeds, 15.9% had the haplotype common to the domesticated cultivars (haplotype 2). In the GNM, 6 out of 14 positive seeds for Cry1Ab/Ac have haplotype 2. In the GSM, three out of five individuals positive for CP4-EPSPS had this haplotype. In the NPM, two seeds positive for Cry1Ab/1Ac and Cry2A shared this haplotype. In the SPM, none of the positive seeds for recombinant proteins had this haplotype.

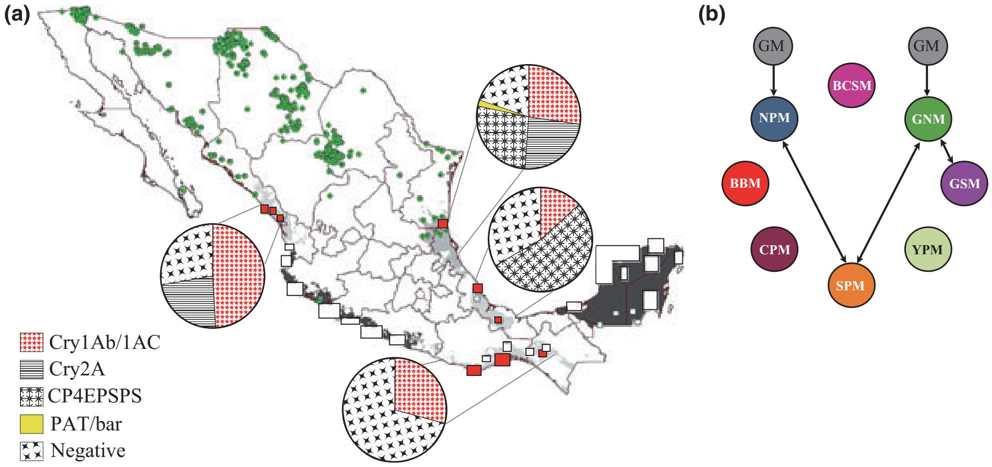

Fig. 4 Contemporary gene flow among cotton metapopulations as inferred by transgene presence. (a) Map of Mexico showing the regions where GM cotton cultivars have been approved for planting, as well as wild cotton metapopulations and populations positive for recombinant protein presence. GM cotton cultivation sites are plotted as green circles; metapopulations without recombinant proteins (BCSM, BBM, CPM, and YPM) are coloured in dark grey; metapopulations with recombinant proteins (NPM, SPM, GNM and GSM) are in pale grey; wild cotton populations with transgene presence are plotted as red squares while populations without transgenes are depicted as white squares. Pie charts with the frequency of particular recombinant proteins are set aside each transgene-harbouring metapopulation. (b) Diagram showing possible venues of present gene flow between GM cultivars and some wild cotton metapopulations. Arrows show the probable trajectories of transgene flow.

Discussion

In this study, we have shown that long distance gene flow has taken place among *G. hirsutum* wild populations, both historically and recently. Evidence from the population genetic analyses and the metapopulation scheme suggests that geographical barriers can hinder population structuring, but not sufficiently to suppress migration among metapopulations.

Cotton metapopulations: current distribution, metapopulation dynamics, and changes in land use

In this work, we propose the existence of eight distinct wild cotton metapopulations in Mexico. While the standards utilized in this work to define metapopulations are qualitative, they are consistent with the delimitation criteria put forward by other scholars (Hanski 1998; Freckleton & Watkinson 2003; Honnay *et al.* 2009). In line with metapopulation theory, we found that 34% of the historically characterized *G. hirsutum* populations continue to dwell in their original geographic zones. Local extinction and recolonization was also observed in 68 of 171 collection points surveyed for which at least two visits were performed during this work.

This dynamic turnover could be favoured by the fact that 55% of surveyed wild populations live in disturbed areas. This suggests, based on Fryxell's and others' previous assessments of wild cotton populations, that a process of habitat alteration due to human and abiotic perturbations (changes in land use, as well as hurricanes and tropical storms) has taken place. These phenomena have shaped the species' habitat along the Pacific and Gulf of Mexico coastal lines. The survival of these populations in disturbed areas is probably related to the ability of this species to grow well in places with low plant cover and high solar exposure, as well as having a perennial habit, being sexually mature during the first year of life, having populations composed of plants at different life stages, and presenting long distance seed dispersal. Nevertheless, while habitat perturbations have not affected all cotton populations, they could drive a significant number of them to extinction, especially in a scenario where extreme changes in land use would hinder recolonization. This could be the case for the coastal region of the Gulf of Mexico (GNM and GSM), which has been subjected to land use changes due to promotion of agriculture and cattle grazing areas (GNM) or to the establishment of hotel resorts that deplete coastal dunes (GSM). This probably accounts for the smaller number of wild cotton populations documented in this study for that part of the country.

While the dynamics currently affecting metapopulation structuring are probably significantly influenced by human activity, the structure unveiled in this work can only be explained in evolutionary time. The modelling of the ecological niche for *G. hirsutum* populations in Mexico was also based on data obtained from actual wild cotton populations. This confers a more precise

Table 2 Linear distances between GM cotton release-sites and wild cotton metapopulations

Cotton growing region	Minimum distance between GM cotton plot and a wild cotton metapopulation (km)	Number of granted permits (1996–2008)
Tamaulipas and Sinaloa	0	85
Tamaulipas, Sinaloa and South Sonora	1–100	152
	101–200	42
	201–300	56
Comarca Lagunera	301–400	919
	401–500	1200
South Chihuahua	501–600	378
	601–700	274
North Chihuahua	701–800	1375
North Sonora	801–900	210
	901–1000	210
Mexicali, Baja California Norte	1001–1100	1084

input for distribution inference algorithms such as GARP, than was obtained by previous studies where this distribution was inferred using data from cultivated cotton (Rogers *et al.* 2007).

Ancestral gene flow among cotton metapopulations and cotton cultivars

We evaluated historical gene flow using chloroplast microsatellites (maternally inherited alleles) to detect historical gene flow through seed migration. Our data indeed suggest long distance seed migration that is consistent with previous suggestions of the potential for seed dispersal through marine currents, given the viability of seeds subjected to prolonged incubation periods in salt water (Stephens 1958). This finding is consistent with the signature of molecular markers during chromosomal speciation (Wendel 1989; Wendel & Albert 1992; Andersson & de Vicente 2010). Interestingly, when assessing recent long distance gene flow through transgene presence in wild *G. hirsutum* populations, we find high migration rates (m = 66/270 = 0.24), but this does not seem to be due to seed migration, since only 15.9% of the plants that were positive for transgene presence have the haplotype common to domesticated cotton used to generate transgenic lines (haplotype 2). This observation could imply low seed migration out of GM fields. Nevertheless, once a single or a few transgenic individuals are dispersed into particular wild populations, they produce pollen that may fertilize local wild plants. Since transgenes are inserted within the nuclear genome, they can be dispersed both via pollen or seed.

As cotton was domesticated centuries ago, ancient gene flow between domesticated cultivars and its wild relatives could probably have occurred historically via seed dispersal, favoured by human activities and environmental phenomena. Thus, some of the genetic patterns observed could be the product of these types of ancestral events. Nevertheless, we assumed that the observed genetic structure is affected by historic gene flow events among cultivated and wild cotton and we repeated the haplotype analysis, this time eliminating haplotype 2 (the only haplotype in cultivated specimens). We did not find significant changes with respect to the structure reported here (data not shown).

The haplotype network that we have put forward has helped us to distinguish ancient polymorphisms from recent gene flow events. Furthermore, these approximations have been complemented by the estimation of recent gene flow using transgenes as markers in extant wild cotton populations (Fig. 4).

Transgenes in wild cotton metapopulations

Fifteen years after the introduction of GM cotton cultivars into Mexico, we have documented the presence of recombinant proteins in wild cotton populations at its CCO (see Fig. 4a). We assayed recombinant protein activity using ELISA kits available in Mexico. These enabled us to detect 18 out of 21 approved events (CERA 2010) among individuals of wild cotton populations. The remaining undetectable events (3) have been scarcely sown. The traits that have been introduced, alone or in different combinations, into currently sown cotton varieties through genetic engineering include Lepidoptera resistance (*Cry1Ab/Ac*, *Cry2Ac*, *Cry1F* and *vip3A*), herbicide tolerance (CP4-EPSPS), and antibiotic resistance (*PAT/Bar*, *nptII* and *aph4*; Traxler & Godoy-Avila 2004).

The combinations of recombinant proteins detected in this study differ among metapopulations, which suggests that each combination could have been the result of independent and multiple transgene flow events into the Mexican wild cotton populations. This observation is additionally supported by the fact that 84.1% of seeds that tested positive for transgene expression had a haplotype other than the one present in the cultivars (2). Since cotton is assumed to be self pollinated, transgene flow must also have occurred mostly via seed and secondary cross-pollination events (Dyer *et al.* 2009).

The combinations of transgenes found within metapopulations and the possible transgenic events from which they could have originated are as follows: in PNM, plants expressing Cry1Ab/Ac, could have originated from event MON531; in GSM, CP4-EPSPS protein could involve either MON88913 or MON1445/1698; for the GNM and GSM metapopulations, the recombinant protein combinations found -Cry1Ab/Ac and CP4-EPSPS- suggest that the most likely transgenic event could be MON531 × MON1445. For GNM, the Cry1Ab/Ac, Cry2Ac, and CP4-EPSPS proteins could originate from MON15985 × MON1445 or MON88913. These events were approved for planting in Mexico between 1996 and 2003. In the case of seeds positive for transgenes that harbour haplotype 2 and have transgene combinations consistent with a commercial transgenic variety (15.9%), we could be detecting feral GM cotton plants that have dispersed into suitable habitats, but, given the environmental conditions, do not grow to resemble their cultivated counterparts.

In contrast, we found some transgene combinations that cannot be explained as primary gene flow events, given the transgene combinations present in the currently available GM cotton lines. This is the case of a seed from GNM that expresses all four recombinant proteins assayed. This finding suggests that recurrent gene flow events and gene stacking could already have occurred in this metapopulation. In contrast, some seeds from NPM and SPM only expressed the Cry2Ac protein, which is not contained individually in any commercial event. This phenomenon could involve independent segregation of transgenes from some lines and a later introgression into wild cotton. Alternatively, it could represent transcriptional or post-transcriptional gene silencing of either the CP4-EPSPS or Cry1Ab gene/transcripts that are present in all commercially available lines expressing Cry2Ac. In order to distinguish between these hypotheses, DNA-based analyses should be undertaken using transgene specific primers, both from the DNA of the seed and that of the mother plant.

Detected transgenes were aggregated in space ($p = 0.001$). This type of distribution could be favoured by the dynamics of plant metapopulations. In the particular case of wild cotton, populations where recolonization has taken place have few plants, and thus can be subject to a genetic bottleneck and to genetic drift. Transgene frequencies and spatial patterns documented here also suggest that transgene introduction is relatively recent and has not been fixed in all metapopulations. These findings could imply that these new alleles do not confer a high selective advantage. In the particular metapopulations where not all populations are positive for transgene presence (GSM and PSM), but are in close proximity to some of the positive populations, three hypotheses would explain the intermixing of positive and negative populations. Firstly, negative populations are more recent than positive ones and are the product of colonization/recolonization from wild (non-transgenic) seed. Alternatively, these populations did have transgenes, but the transgenes were eliminated by genetic drift or selection. Finally, transgenes in these populations may exist but are silenced or have not reached some populations simply due to random events.

Our transgene data confirm that long-distance gene flow is preeminent in wild cotton at its CCO. Given present day management practices, some means of seed movement at long distances include the accidental dispersal of cotton seed intended for animal feed. We observed this happening in trucks from the USA to the centre-south of Mexico. This phenomenon takes place because seeds that are separated from their fibre are later sold as animal feed without being previously mashed into a 'cake'. This venue for GM seed dispersal could very well be occurring for GM seed processed in Mexico, because very little attention is paid to the disposal of this seed once the fibre has been removed. Given the documented patterns, future studies should address the possible scenarios to be expected in terms of transgene flow and accumulation, as well as the consequences these may have for wild cotton conservation at its CCO, as has been documented for maize in Mexico (Dyer & Taylor 2008; Piñeyro-Nelson *et al.* 2009).

In this study, we found no correlation between transgene presence and loss of genetic diversity. Nevertheless, in order to explore whether the presence of transgenes could have consequences in wild cotton populations, sustained and long-term analyses should be pursued. In particular, biomonitoring studies that assess the consequences of both transgene and foreign germplasm introduction into wild metapopulations of *G. hirsutum*, should be undertaken (see, for example, Meirmans *et al.* 2009).

This study confirms that ELISA-based analyses are useful when assessing the presence of transgenes in wild cotton metapopulations. Nevertheless, future studies should also consider DNA-based detection methods to corroborate our findings, as well as to determine the specific events involved. This multiple-technique approach has been suggested in other studies dealing with transgene detection at CCO (Serratos-Hernández *et al.* 2007; Piñeyro-Nelson *et al.* 2009).

Lastly, gene flow from cultivated cotton can put the wild germplasm of several *Gossypium* species at risk. Evidence from previous investigations suggests that *G. tomentosum* (in Hawaii), *G. mustelium* (in Brazil) and *G. darwinii* (in Galapagos) are at risk of extinction as a result of hybridization with domesticated tetraploid cotton (Ellstrand 2003; Andersson & de Vicente 2010). In some cases, interspecific hybrids (*G. hirsutum* × *G. barbadense*) may act as genetic bridges for gene transfer from domesticated cotton to other wild relatives (*G. darwinii*; Ellstrand 2003; Andersson & de Vicente 2010). As a consequence, conservation programs should include all *Gossypium* tetraploid species.

Conclusions

The interplay of historical long distance gene flow and geographic barriers in Mexico has shaped the genetic structure of extant populations of *G. hirsutum*. Extinction and recolonization events in particular populations have hindered genetic homogenization among metapopulations.

Potential distribution analyses and molecular markers independently show the existence of eight metapopulations. We were able to record intense dynamics of recent local extinctions and colonizations that go back to the collections made by Paul Fryxell. In spite of their integrity, these metapopulations are connected through long distance migration events. In particular, through the assessment of transgene presence, we were able to detect recent gene flow, which supports the connectivity of these metapopulations. This scenario of long distance colonization, the existence of metapopulations, and the presence of transgenes at its CCO calls for conservation efforts both *in situ* and *ex situ*. These types of endeavours rely upon preservation of the habitat currently occupied by wild cotton plants or on opening up of new habitats for wild cotton colonization. The metapopulation perspective must be kept in mind (Meirmans *et al*. 2003), as 'metapopulation persistence relies on the existence of a certain amount of suitable but currently unoccupied habitat' (Freckleton & Watkinson 2002, 2003). Coastal dunes appear to be particularly important areas in this respect. In addition, demographic studies of wild cotton populations, documentation of spatio-temporal patterns of seed and pollen dispersal, and rates of cross-pollination among wild individuals should also be the basis for guiding these conservation efforts.

Acknowledgements

The authors thank Victoria Sork, René Cerritos, José Casar, Lluvia Flores, Valeria Alavez and three anonymous reviewers for useful comments on earlier versions of this manuscript; Rafael Lira, Alejandro Casas, Judith Márquez-Guzmán, Luis Felipe Jiménez, and Francisca Acevedo gave useful advice for this work. AW and APN would like to thank Alejandra Vázquez-Lobo, Rodolfo Salas, Argelia Cuenca, Lev Jardón, Diego Ortega, Adriana Caballero and Luz María Martínez for their help in field-work and in various laboratory tasks. PhD scholarships from Consejo Nacional de Ciencia y Tecnología and Universidad Nacional Autónoma de México were granted to AW and APN. This work is part of the PhD. dissertation of AW in Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México. Research was done with support from CONABIO, through grants GEF-CIBIOGEM 00013572, PNUD-CIBIOGEM 00051868 to AW and DP. ERAB and APN research was financed by grants UC-Mexus ECO-IE359 and 'Red Temática de Investigación CONACYT: Complejidad, Ciencia y Sociedad' (124909).

References

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modeling*, **162**, 211–232.

Andersson MS, de Vicente CM (2010) *Gene Flow Between Crops and Their Wild Relatives*, 564 pp. Johns Hopkins University, Baltimore.

Applequist WL, Cronn R, Wendel JF (2001) Comparative development of fiber in wild and cultivated cotton. *Evolution & Development*, **3**, 3–17.

Arnaud J-F, Fénart S, Godé C, Deledicque S, Touzet P, Cuguen J (2009) Fine-scale geographical structure of genetic diversity in inland wild beet populations. *Molecular Ecology*, **18**, 3201–3215.

Baltazar BM, Sanchez-Gonzalez JD, Cruz-Larios L, Schoper JB (2005) Pollination between maize and teosinte: an important determinant of gene flow in Mexico. *Theoretical and Applied Genetics*, **110**, 519–526.

Bartsch D, Lehnen M, Clegg J, Pohl-Orf M, Schuphan I, Ellstrand NC (1999) Impact of gene flow from cultivated beet on genetic diversity of wild sea beet populations. *Molecular Ecology*, **8**, 1733–1741.

Beasley JO (1940) Hybridization of American 26 chromosome and Asiatic 13 chromosome species of *Gossypium*. *Journal of Agricultural Research*, **69**, 175–181.

Beasley JO (1942) Meiotic chromosome behavior in species hybrids, haploids, and induced polyploids of *Gossypium*. *Genetics*, **27**, 25–54.

Beekman M, Ratnieks LF (2000) Long-range foraging by the honey-bee, *Apis mellifera* L. *Functional Ecology*, **14**, 490–496.

Broquet T, Petit EJ (2009) Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 193–216.

Brubaker CL, Wendel JF (1994) Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). *American Journal of Botany*, **81**, 1309–1326.

CERA (2010) *GM Crop Database*. Center for Environmental Risk Assessment (CERA), ILSI Research Foundation, Washington DC. http://cera-gmc.org/index.php?action=gm_crop_database (Accessed on February 1, 2011).

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1660.

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (2003) *México: Imagen desde el espacio*. CONABIO, México, DF.

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modeling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Dyer RJ (2009) GeneticStudio: a suite of programs for the spatial analysis of genetic marker data. *Molecular Ecology Resources*, **9**, 110–113.

Dyer RJ, Nason JD (2004) Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, **13**, 1713–1727.

Dyer JA, Taylor JE (2008) A crop population perspective on maize seed systems in Mexico. *Proceedings of the National Academy of Sciences, USA*, **105**, 470–475.

Dyer G, Serratos-Hernández A, Perales H *et al.* (2009) Dispersal of transgenes through maize seed systems in Mexico. *PLoS ONE*, **4**(5), e5734.

Ebert D, Haag D, Kirkpatrick M, Riek M, Hottinger JW, Pajunen VI (2002) A selective advantage to immigrant genes in a Daphnia metapopulation. *Science*, **295**, 485–487.

Ehrlich R, Raven PH (1969) Differentiation of populations. *Science*, **165**, 1228–1232.

Ellstrand NC (2003) *Dangerous liaisons? When Cultivated Plants Mate With Their Wild Relatives*, 244 pp. Johns Hopkins University, Baltimore.

Ellstrand NC, Garner LC, Hegde S, Guardagnuolo R, Blancas L (2007) Spontaneous hybridization between maize and teosinte. *Journal of Heredity*, **98**, 183–187.

Excoffier L, Lischer HEL (2010) Arlequin suite v3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

FAOSTAT (2009) *Statistical Databases*. Food and Agriculture Organization of the United Nations, Rome, Italy. http://faostat.fao.org (accessed on February 1, 2011).

Fénart S, Austerlitz F, Cuguen J, Arnaud J-F (2007) Long distance pollen-mediated gene flow at a landscape level: the weed beet as a case study. *Molecular Ecology*, **16**, 3801–3813.

Freckleton RP, Watkinson A (2002) Large-scale spatial dynamics of plants: metapopulations, regional ensembles and patchy populations. *Ecology*, **90**, 419–434.

Freckleton RP, Watkinson AR (2003) Are all plant populations metapopulations? *Journal of Ecology*, **91**, 321–324.

Fryxell PA (1979) *The Natural History of the Cotton Tribe*, 245 pp. Texas A & M University Press, College Station/London.

Hanski I (1998) Metapopulation dynamics. *Nature*, **396**, 41–49.

Hanski I, Gaggiotti O (2004) *Ecology, Genetics and Evolution of Metapopulations*. Elsevier Academic Press, London, UK.

van Heerwaarden J, Van Eeuwijk FA, Ross-Ibarra J (2009) Genetic diversity in a crop metapopulation. *Heredity*, **104**, 28–39.

Honnay O, Jacquemyn H, Van Looy K, Vandepitte K, Breyne P (2009) Temporal and spatial genetic variation in a metapopulation of the annual *Erysimum cheiranthoides* on stony river banks. *Journal of Ecology*, **97**, 131–141.

Jørgensen RB, Andersen B (1994) Spontaneous hybridization between oilseed rape (*Brassica napus*) and weedy *B. campestris* (Brassicaceae). *American Journal of Botany*, **81**, 1620–1626.

Kalinowsky ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics*, **5**, 539–543.

Levins R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America*, **15**, 237–240.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences, USA*, **103**(25), 9578–9583.

Manel S, Schwartz K, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**, 189–197.

McGregor SE (1976) *Insect Pollination of Cultivated Crop Plants*. US Department of Agriculture, *Agriculture Handbook*, 496, 411.

Meirmans PG, Bousquet J, Isabel N (2009) A metapopulation model for the introgression from genetically modified plants into their wild relatives. *Evolutionary Applications*, **2**, 160–171. doi:10.1111/j.1752-4571.2008.00050.x.

Paetkau D, Slade R, Burdens M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology*, **13**, 55–65.

Palstra FP, O'connell MF, Ruzzante DE (2007) Population structure and gene flow reversals in Atlantic salmon (*Salmo salar*) over contemporary and long-term temporal scales: effects of population size and life history. *Molecular Ecology*, **16**, 4504–4522.

Petit RJ, Mousadirk A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.

Piñeyro-Nelson A, Van Heerwaarden J, Perales HR *et al.* (2009) Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. *Molecular Ecology*, **18**, 750–761.

Reddy OU, Pepper AE, Abdurakmonov I *et al.* (2001) New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research. *Journal Cotton Science*, **5**, 103–113.

Rogers DJ, Reid RE, Rogers JJ, Addison SJ (2007) Prediction of the naturalization potential and weediness risk of transgenic cotton in Australia. *Agriculture Ecosystems and Environment*, **119**, 177–189.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.

SAGARPA (2010) http://www.sagarpa.gob.mx (Accessed on February 1, 2011).

Sahoo L, Schmidt JJ, Pedersen JF, Lee DJ, Lindquist JL (2010) Growth and fitness components of wild x cultivated *Sorghum bicolor* (Poaceae) hybrids in Nebraska. *American Journal Botany*, **97**, 1610–1617.

Sasu MA, Ferrari MJ, Du D, Winsor JA, Stephenson AG (2009) Indirect costs of a nontarget pathogen mitigate the direct benefits of a virus-resistant transgene in wild *Cucurbita*. *Proceedings of the National Academy of Sciences, USA*, **106**, 19067–19071.

Scachetti-Pereira R (2001) *Desktop GARP*. http://www.nhm.ku.edu/desktopgarp/index.html (Accessed on July 2010).

Seelanan T, Schnabel A, Wendel JF (1997) Congruence and consensus in the cotton tribe (Malvaceae). *Systematic Botany*, **22**, 259–290.

Serratos-Hernández JA, Gómez-Olivares JL, Salinas-Arreortua N, Buendía-Rodríguez E, Islas-Gutiérrez F, de-Ita A (2007) Transgenic proteins in maize in the soil conservation area of Federal District, México. *Frontiers in Ecology and the Environment*, **5**, 247–252.

Simpson D (1954) Natural cross-pollination in cotton. *US Department* of *Agriculture Technical Bulletin*, 1094.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.

Smith C, Stephens S (1971) Critical identification of Mexican archaeological cotton remains. *Economic Botany*, **25**, 160–168.

Snow AA, Andersen B, Jorgensen RB (1999) Costs of transgenic herbicide resistance introgressed from *Brassica napus* into weedy *B. rapa*. *Molecular Ecology*, **8**, 605–615.

Stephens SG (1958) Factors affecting seed dispersal in *Gossypium* and their possible evolutionary significance. *North Carolina Agricultural Experiment Station Technical Bulletin*, No. 131. 32 pp.

Stephens SG (1966) The potentiality for long range oceanic dispersal of cotton seeds. *The American Naturalist*, **100**, 199–210.

Stephens SG, Finkner MD (1953) Natural crossing in cotton. *Economic Botany*, **7**, 257–269.

Sul IW, Korban SS (1996) A highly efficient method for isolating genomic DNA from plant tissues. *Plant Tissue Culture Biotechnology*, **2**, 113–116.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics*, **24**, 2498–2504.

Templeton AR, Sing CF (1993) 'A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Traxler G, Godoy-Avila S (2004) Transgenic cotton in Mexico. *AgBioForum*, **7**, 57–62.

Viard F, Arnaud J-F, Delescluse M, Cuguen J (2004) Tracing back seed and pollen flow within the crop-wild *Beta vulgaris* complex: genetic distinctiveness Vs. hot spots of hybridization over a regional scale. *Molecular Ecology*, **13**, 1357–1364.

Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledenous angiosperms. *Genome*, **42**, 9–19.

Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Science of USA*, **86**, 4132–4136.

Wendel JF, Albert VA (1992) Phylogenetics of the cotton genus (*Gossypium* L.): character-state weighted parsimony analysis of chloroplast DNA restriction site data and its systematic and biogeographic implications. *Systematic Botany*, **17**, 115–143.

Wendel JF, Brubaker CL, Álvarez I, Cronn RC, Stewart J McD (2009) Evolution and natural history of the cotton genus. In: *Genetics and Genomics of cotton, Plant Genetics and Genomics: Crops and Models 3* (ed. Paterson AH), pp. 1–20. Springer, New York.

Wendel JF, Brubaker CL, Seelanan T (2010) The origin and evolution of *Gossypium*, Chapter 1. In: *Physiology of Cotton* (eds Stewart JM, Oosterhuis D, Heitholt JJ, Mauney JR), pp. 3–22. Springer, Netherlands.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: F_{ST} doesn't equal 1/(4Nm+1). *Heredity*, **82**, 117–125.

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, **16**, 120–127.

A.W. is interested in the evolution, applied population genetics of plants, as well as the study of the centers of origin and diversification of cultivated plants. A.P.-N. is interested in molecular genetics and plant evolutionary development. J.A. is interested in modeling the distribution of species and in the analysis of spatial information. A.G. is interested in molecular detection and quantification of GM sequences as well as heterologous proteins. E.R.Á.-B. is interested in genetics, evolutionary development and plant conservation as well as biomathematics. D.P. is interested in the fields of population genetics and phylogeography. He is currently involved in the study of Mexican pines.

Data accessibility

Microsatellite data available in DRYAD: doi:10.5061/dryad.rd8fn

Supporting information

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

Table S1 Description of DNA extraction procedure and PCR conditions

Table S2 Pairwise comparison matrix of genetic distance and genetic diversity among metapopulations

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.