Landscape connectivity of a noxious invasive weed: promises and challenges of landscape genomics for knowledge-based weed management?

Diego F. Alvarado-Serrano (corresponding author)
Department of Ecology and Evolutionary Biology
University of Michigan
E-mail: dalvarad@umich.edu

Megan Van Etten
Department of Ecology and Evolutionary Biology
University of Michigan
E-mail: mvanette@umich.edu

Shu-Mei Chang
Department of Ecology and Evolutionary Biology
University of Michigan
E-mail: smchang@uga.edu

Regina S. Baucom
Department of Ecology and Evolutionary Biology
University of Michigan
E-mail: rsbaucom@umich.edu
Landscape connectivity of a noxious invasive weed: promises and challenges of landscape genomics for knowledge-based weed management?

ABSTRACT

Examining how the landscape may influence gene flow is at the forefront of understanding population differentiation and adaptation. Such understanding is crucial in light of ongoing environmental changes and the elevated risk of ecosystems alteration. In particular, knowledge of how humans may influence the structure of populations is imperative to allow for informed decisions in management and conservation. Here we characterize the population genetic structure of *Ipomoea purpurea*, a noxious invasive weed, and assess the interaction between natural and human-driven landscapes on genetic differentiation. By combining rigorous statistical analyses and different molecular markers (nuclear microsatellites and a genome-wide panel of SNPs), we detect both common and marker-specific patterns of genetic connectivity and identify human population density as an important predictor of pairwise population differentiation, suggesting that the agricultural and/or horticultural trade may be involved in maintaining some level of connectivity across distant agricultural fields. Climatic variation appears as an additional predictor. We discuss the implications of these results and the approach we followed in the context of understanding agricultural weed and invasive species’ connectivity, as well as the challenges and promises of current landscape genomics research for knowledge-based weed management.

Keywords: agricultural weeds, human-aided migration, landscape genetics, morning glory, population structure
INTRODUCTION

Elucidating routes and levels of migration between subpopulations of a species is essential to understand the interplay between gene flow, adaptation, genetic drift, and selection, and hence the forces that shape its evolutionary trajectory (Barrowclough, 1980; Slatkin, 1985). Landscape features—such as rivers, mountain ranges, crop fields, and urban areas—can impact levels of gene flow between populations by determining dispersal rates and routes (Cushman, McKelvey, Hayden, & Schwartz, 2006; McRae, 2006) as well as by influencing the likelihood of successful establishment of immigrants (Nosil, Egan, & Funk, 2008; Sexton, Hangartner, & Hoffmann, 2014; Wang & Bradburd, 2014). Landscape features can also indirectly condition the effect of gene flow through its effect on local effective population sizes since the actual role that migration plays in the evolution of a species is driven by the fraction of the local population size that correspond to immigrants (Slatkin, 1985; Wright, 1949). Consequently, the landscape, loosely defined as an area with spatially variable biotic and abiotic factors (Holderegger, Buehler, Gugerli, & Manel, 2010), creates the stage for spatially variable levels of effective gene flow among populations, conditioned by species’ specific physiological tolerances and behavioral preferences (Clobert, Baguette, Benton, Bullock, & Ducatez, 2012). In this way, the landscape plays a pivotal role in the evolution of species.

In contrast to species that depend almost exclusively on natural dispersal agents, species in heavily human-dominated ecosystems may exploit human activities to maintain gene flow among populations and expand their ranges (Everman & Klawinski, 2013; Fountain, Duvaux, Horsburgh, Reinhardt, & Butlin, 2014). Such species may be capable of maintaining population connectivity over vast geographic ranges (Trakhtenbrot, Nathan, Perry, & Richardson, 2005) by overcoming landscape features that would otherwise represent natural barriers and reach dispersal
distances that could be orders of magnitude greater than those attained under natural agents or do it under much smaller time frames (Mack & Lonsdale, 2001; Ricciardi, 2007). In this way, by facilitating dispersal humans have the potential to: i) condition the balance between drift and selection (Lenormand, 2002; Slatkin, 1985), ii) introduce relevant genetic variation to local populations (Kolbe et al., 2004), iii) prevent local extinction or favor recolonization (Fountain et al., 2014), and alter the overall genetic constitution of populations (Bataille, Cunningham, Cruz, Cedeño, & Goodman, 2011). Human-aided migration—intentional or unintentional—is particularly prevalent in plants (Auffret & Cousins, 2013; Hodkinson, Thompson, Journal, & Dec, 2007; Wichmann et al., 2009), where it has had major impacts on the distribution of species and stability of communities (Simberloff, 2013 and references therein). Yet, the open question remains: how might the interaction between human-dominated and more natural landscapes affect population connectivity in plant populations, especially in human-exploiter species?

A particularly amenable system to study the interaction between natural and human-aided dispersal comes from agricultural weed populations. Agricultural weeds face a highly dynamic landscape characterized by frequent spatial rearrangements (expansion of agricultural front, increased fragmentation) and a constantly changing environment (crop rotation, agricultural chemical use, climatic abnormalities) (Meehan, Werling, Landis, & Gratton, 2011; Menchari, Délye, & Le Corre, 2007) that certainly impact their opportunities for survival and local adaptation through its effect on population connectivity (Margosian, Garrett, Hutchinson, & With, 2009). At the same time, natural features such as climate, soil type, and topography are expected to play a significant role in structuring populations provided intrinsic physiological requirements and species-specific traits (Cimalová & Lososová, 2009; Navas, 2012). Under these conditions, human-aided migration is expected to be critical for weeds’ success (Epperson 

Clegg, 1986), but knowledge on how or if weedy plant populations are able to maintain connectivity through a complex landscape matrix of croplands, grasslands, natural and urban areas is limited. Addressing this limitation should not only improve our understanding of the underlying processes governing weeds’ population structure, but should also offer practical tools to deal with this ever-growing agricultural problem that impose severe economic costs (on the order of 33B USD per year in US agriculture alone; Pimentel, Zuniga, & Morrison, 2005).

As a first step into investigating the interplay between natural factors and human activities on structuring genetic diversity in weed populations, we estimate the intensity and extent of migration from genetic data and, under the preliminary simplifying assumption of evolutionary equilibrium (Marko & Hart, 2011), evaluate how multiple landscape features influence genetic connectivity of a noxious agricultural weed, Ipomoea purpurea, using two different sets of molecular markers (nuclear microsatellites and a genome-wide panel of SNPs). Specifically, we ask the following questions: 1) Which natural and/or human-influenced landscape features—soils, elevation, climate, landcover, crop types, human population density—promote or constrain genetic connectivity between populations of this agricultural weed? and 2) what additional insights can we gain from a broader representation of the genome than traditionally used in landscape genetics studies (typically microsatellites and organelle DNA)? By considering the possible interactions between natural and human effects on migration, the answers to these questions offer deeper knowledge of the interaction between human activities, landscape features, and population structure of noxious weeds and hence contribute to improve effective management and control of these damaging plants. More generally, these answers contribute to deepen our understanding of the interaction between environmental setting and population differentiation, adaptation, and persistence (Taylor, Fahrig, Henein, & Merriam, 1993).
MATERIALS AND METHODS

Study system

*Ipomoea purpurea*, the common morning glory, is an agricultural weed evolving under the influence of human-driven and natural landscape factors. This species is a noxious weed, of horticultural value (Defelice, 2001; Fang et al., 2013), with a widespread distribution that includes highly heterogeneous landscapes in the Eastern, South- and Mid-western regions of the United States (Culpepper, 2006; Webster & Nichols, 2012). It is a self-compatible annual bumblebee-pollinated vine, with heavy seeds, and is found primarily in agricultural fields and disturbed areas (Baucom & Mauricio, 2008; Tiffin & Rausher, 1999), as well as cultivated flower gardens and yards (Defelice, 2001). While most details on the history of *I. purpurea* remain unknown, it is hypothesized that *I. purpurea* originated in Central America, from where it was taken to Spain to be grown in monasteries as an ornamental during the XVI century (Defelice, 2001; Fang et al., 2013). From there, it is hypothesized that its cultivation expanded to other European countries, including England, and later to North America (Defelice, 2001; Fang et al., 2013). By the early XVIII century it became a popular plant in gardens in the United States, but also a known weed (Defelice, 2001). Since then, little is know about the demographic history of this species in the United States, other than the fact that gene flow has probably been maintained over time at least among populations in relatively close geographic proximity (Kuester, Chang, & Baucom, 2015). What is clear is that since its introduction its history has been tightly linked to human activities, making it suitable to assess the impact of natural and anthropogenic landscapes on structuring genetic variance.
Ipomoea purpurea is currently one of the most problematic agricultural weeds (Webster & Nichols, 2012) and is capable of infestations leading to substantial decline in crop (closely related Ipomoea species cause declines of up to 80% of crop yield; Rogers, Murray, Verhalen, & Claypool, 1996). This species exhibits resistance to the commonly used herbicide glyphosate (Baucom & Mauricio, 2004, 2008), although the exact resistance level varies widely among populations of this species (Kuester et al., 2015). This species is also a major concern for conservation given its naturalization in multiple regions throughout the world and its aggressiveness as an invasive (Chaney & Baucom, 2012; Fang et al., 2013). Hence, unraveling the population structure of this species and how it is affected by the landscape should not only improve our understanding of basic evolutionary processes, but should also inform practical decisions for its management and control (e.g., Is herbicide resistance better controlled by avoiding the spread of resistance genotypes or by local management of moderately isolated populations?).

**Data compilation**

To capture the plausible effect of both natural and disturbed landscapes on structuring genetic diversity in I. purpurea, we compiled a diverse set of GIS data for the continental US from a variety of sources (Table S1). These data encapsulate human activities (human population density, landcover, planted crops, and roads) as well as the geographical setting of I. purpurea (elevation, climate—19 variables summarizing central tendencies and variability patterns in temperature and precipitation, soil—8 variables summarizing the texture, pH, and organic and inorganic content of the top 20cm of soil). We focused on both sets of data because of the possible interaction between natural and human effects, which may lead to incorrect inferences if not accounted for (e.g., spurious associations due to spatial correlation between crops distribution...
and climate; Eberhardt & Teal, 2013). We first processed all these data into landscape layers at a common spatial resolution of 10km² and a common spatial extent around the US states with available samples (Fig. 1). These spatial resolution and extent were chosen to maintain a practical balance between scale and analytical manageability given available computational resources. To reduce dimensionality, we opted to perform two separate Principal Component Analyses (PCAs) on the 19 climatic and 8 soil layers, respectively. For all subsequent analyses we kept the resulting first two principal components of each of these analyses, which accounted for over 78% of the variance in each case, and primarily summarized temperature temporal gradients and precipitation seasonality, and soils’ pH, sandiness, and grain size, respectively (Table S2).

With the objective of estimating the genetic connectivity of populations of *I. purpurea*, we compiled genetic data on an extensive panel of 15 previously optimized microsatellite loci (Molecular Ecology Resources Primer Development Consortium 2013), which quality has been verified by looking for scoring errors (Kuester et al., 2015) using Micro-Checker (Van Oosterhout, Hutchinson, Wills, & Shipley, 2004). These data encompass a total of 597 individuals from 31 localities (with a minimum of 8 individuals per locality) (Fig. 1; Table S3). All individuals were collected in 2012 from farms across the range of *I. purpurea* in the United States (Kuester et al., 2015). In addition, to obtain a more comprehensive representation of the genome of *I. purpurea* and assess the robustness of results in light of coalescent and mutational variance (Nielsen & Slatkin, 2013), we generated a Next Generation Sequencing (NGS) dataset from an additional set of individuals (from 6 localities represented in the SSR dataset, plus 2 additional localities in close geographic proximity to localities in the SSR dataset; Fig. 1).
To generate the NGS dataset, we constructed genome-wide Genotype By Sequencing (GBS) libraries for 80 individuals sampled across the 8 localities. The GBS library was developed using 7ng of genomic DNA, extracted from leaf or cotyledon tissue, using SNPsaurus’ (Oregon, USA) nextRAD technology. This technology uses a selective PCR primer to amplify consistent genomic loci among individuals. Similarly to RAD-Seq sequences (Rowe, Renaut, & Guggisberg, 2011) in which the DNA flanking a restriction enzyme cut site is selected for amplification, nextRAD amplifies sequences that correspond to the DNA downstream of a short selective priming site. Samples were first fragmented and then ligated to short adapter and barcode sequences using a partial Nextera reaction (Illumina; California, USA) before being amplified using Phusion® Hot Start Flex DNA Polymerase (New England Biolabs; Massachusetts, USA). The 80 dual-barcoded PCR-amplified samples were pooled and the resulting libraries were purified using AMPure XP beads (Agencourt Bioscience Corporation; Massachusetts, USA) at 0.7x. The purified library was then size selected to 350-800 base pairs and sequenced using two runs of an Illumina NextSeq500 sequencer (Genomics Core Facility, University of Oregon).

The resulting sequences were analytically processed using the SNPsaurus nextRAD pipeline (SNPsaurus, Oregon, USA; Siliceo-Cantero, García, Reynolds, Pacheco, & Lister, 2016). Specifically, reads of 16 randomly selected individuals (of the 80 sequenced) were combined to create a pseudo-reference genome. This was done after removing loci with read counts above 20,000, which presumably corresponded to repetitive genomic material, and loci with read counts below 100, which presumably corresponded to off-target or read errors. The filtered reads were aligned to each other using BBMap (Bushnell, 2016). All parameters were set to default values with the exception of minimum alignment identity, which was set to 0.93 to identify alleles, as it
is a threshold found to work well for non-reference species (SNPsaurus, Oregon, USA). A single read instance was chosen to represent the locus in the pseudo-reference. This resulted in a total of 263,658 loci. All reads from each of the 80 individuals were then aligned to the pseudo-reference using BBMap (Bushnell, 2016) and converted to a vcf genotype table, using Samtools (Li et al., 2009) and bcftools (Li, 2011), after filtering out nucleotides with a quality score of 10 or worse (an empirically informed threshold; SNPsaurus, Oregon, USA). The resulting vcf table was filtered using vcftools (Danecek et al., 2011) for SNPs with a minimum allele frequency of 0.02, a minimum read depth of 5, and a maximum 15% of missing data. We chose this filtering scheme as a balance between accuracy and efficiency and to avoid inadvertent errors associated with our use of a pseudo-reference genome. This resulted in 9774 variable regions. Loci were further filtered using vcftools to exclude loci with less than 5 high quality base-calls and with more than 20% missing data or an average of less than 20 high quality base calls. This resulted in a final panel of 8210 Single Nucleotide Polymorphisms (SNPs) that we used in all subsequent analyses.

Population structure analyses

We first conducted a series of preliminary analyses to characterize the overall genetic structure of *I. purpurea*. All analyses were run separately for the microsatellite (SSR, hereafter) and SNP datasets given their intrinsic differences and distinct geographic coverage (Fig. 1; Table S3). In addition, we repeated all population structure analyses, separate for the SSR and SNP datasets, for the subset of 6 localities with coincident data for both markers (referred as SSRc and SNPc, hereafter) to assess the robustness of results to the difference in geographic coverage.

First, we examined population differentiation by estimating $F_{ST}$ using GenAlex v6.5 (Peakall & Smouse, 2012) (because similar global $F_{ST}$ and $R_{ST}$ estimates were obtained for the SSR...
dataset, we opted to report $F_{ST}$ values only to allow direct comparisons with the SNP dataset). We then estimated contemporary effective population size for each sampled locality in NeEstimator v2 using the excess heterozygous method (Do et al., 2014). We performed this latter analysis to assess the possibility of whether differences in local population size underlie differences in genetic variation (Weckworth et al., 2013) and/or promote asymmetric effective migration rate ($Nm$).

In addition, we assessed population admixture and spatial genetic clustering using TESS (Chen, Durand, Forbes, & François, 2007). TESS was run using the admixture algorithm and a BYM model (Durand, Jay, Gaggiotti, & François, 2009) with 10 runs per K value, and without using geographic weights. The TESS model, with the lowest DIC was chosen as the optimal model (Durand, Chen, & Francois, 2009). K values tested ranged from two to the maximum number of sampled localities. Additionally, following Wang et al. (2009), we complemented these analyses with Analyses of Molecular Variance (AMOVA; Excoffier, Smouse, & Quattro, 1992) run in GenAlex (Peakall & Smouse, 2012) using 9999 permutation replicates. We run these AMOVAs either partitioning the variance into regions based on the spatial genetic clusters previously identified by TESS—to quantify the fraction of the genetic variance explained by these clusters, or leaving it ungrouped (i.e., no regions), for comparison.

Additionally, we investigated population connectivity by estimating levels of recent migration between sampled localities through the identification of individuals of mixed ancestry using BayesAss (Wilson & Rannala, 2003). BayesAss is a program that uses individual multilocus genotypes and a Markov Chain Monte Carlo (MCMC) algorithm to probabilistically distinguish between immigrants and long-term native individuals (Wilson & Rannala, 2003).
ran BayesAss for 6 million generations using default parameter settings, and discarded the first two million generations as burn-in (Dyer, 2009). For each marker dataset, we repeated this analysis three times (for a total of 18 million generations) and combined the results from the three replicates for our final inference. Then, using a posterior probability cut-off of 0.75 we assign individuals’ ancestry. We chose this cut-off value as a minimum credibility score to simultaneously maximize sample size and reliability (stringer thresholds show similar differences between marker sets; results not shown). It is important to note that because of computational limits we had to randomly subsample our set of SNPs to 400 SNPs for this analysis. The same subsampled set was used for the full and reduced (i.e., on the SSRe and SNPe datasets) analyses.

Landscape genetics analyses

After assessing overall population structure of *I. purpurea*, we evaluated the association between landscape features and genetic differentiation based on the full datasets. We limited our analyses to the full datasets because of the robust genetic structure recovered between the full and reduced datasets (see below) and the smaller sample size of the latter datasets, which limits statistical inference power. First, we estimated conditional genetic distances (Dyer, Nason, & Garrick, 2010) using GeneticStudio (Dyer, 2009). Briefly, conditional distances are measures of pairwise genetic distance derived from population networks, constructed based on the degree of genetic similarity between sampled localities (Dyer & Nason, 2004). Because these networks are pruned based on the principle of conditional independence of the total among population genetic covariance (using an edge deviance principle; Magwene, 2001), conditional distances reflect genetic similarity between localities that better capture direct gene flow as opposed to connectivity driven by intervening localities (Dyer, 2015b). The complexity of the associated conditional genetic network was summarized by their vertex connectivity (White & Harary,
2001), whereas the congruence between networks derived from different marker sets was
measured by their structural congruence (a measure of whether the number of congruent edges
between networks is greater than expected by chance) (Dyer, 2009).

Climate, crops, elevation, landcover, population density, roads, and soils landscape layers
(Table S1) were converted into landscape resistance layers by assigning a resistance value to each
landscape feature in these layers to reflect the difficulty that each feature offers to the movement
of gametes or individuals. It is important to note that in contrast to previous studies that typically
rely on expert opinion for resistance assignment, we utilized an unbiased statistical optimization
to avoid the sensitivity of results to subjective resistance assignment (Spear, Balkenhol, Fortin,
McRae, & Scribner, 2010). Specifically, resistance values were optimized through a genetic
algorithm approach (Mitchell, 1996). Briefly, in this search algorithm a population of individuals
with traits encoded by unique combinations of model parameters (resistance assignment
proposals in our case) is allowed to compete with each other based on the fitness associated with
the traits it carries (Peterman, Connette, Semlitsch, & Eggert, 2014). Specifically, in Peterman’s
(2014) implementation of this algorithm, which we followed here, individuals’ fitness is
estimated by the relative quality of a MLPE.lmm model (Maximum Likelihood Population
Effects – Linear Mixture Model) that evaluates the association between pairwise genetic distance
and landscape cumulative resistance between localities, estimated in Circuitscape (Shah &
McRae, 2008). Individuals with parameter settings (i.e., resistance assignments) that result in
better models, as measured by a Deviance Information Criterion (DIC) score, are preferentially
represented in the following generation. Offspring modifications introduced by mutations (i.e.,
small resistance assignment perturbations) allow for exploration of the parameter space. The
algorithm was stopped once 25 generations have passed without significant improvement in fitness.

We implemented Peterman’s (2014) algorithm in R (package ResistanceGA; Peterman, 2014) allowing for the independent optimization of each of our landscape layers. The optimal resistance landscapes identified in this way were then used to run a final univariate MLPE.lmm model to characterize the association between landscape features and conditional genetic distances between localities. Because the roads-association resistance was not recovered as significant for either marker dataset, we dropped this layer for all subsequent analyses. Finally, to identify the simultaneous contribution of natural and human-driven landscape features we ran Multiple Regression on Distance Matrices (MRDM; Legendre, Lapointe, & Casgrain, 1994), which has been identified as one of the best performing methods for evaluating the interplay between landscape features and genetic connectivity (Balkenhol, Waits, & Dezzani, 2009). Before running these MRDM analyses, we standardized all optimized resistance layers to mean of zero and variance of one (Dyer et al., 2010). These final regressions included geographic distance as a null model predictor as well as effective population size and were run in R (package ecodist; Goslee & Urban, 2007) using 10,000 permutations to assess significance. In none of our analyses did we implement a Bonferroni correction for multiple testing because of the overly conservative nature of this correction (Glickman, Rao, & Schultz, 2014; Nakagawa, 2004). Instead we applied a false recovery rate correction (Benjamini & Hochberg, 1995) using the function p.adjust in R (R Core Development Team, 2016).

RESULTS

Population structure
The set of preliminary genetic analyses indicated that *I. purpurea* sampled localities were in no major violation of Hardy-Weinberg equilibrium, as judged by the small difference between expected and observed heterozygosity (mean $H_e = 0.294 \pm 0.014$ and $0.250 \pm 0.001$; mean $H_o = 0.291 \pm 0.009$ and $0.260 \pm 0.001$, respectively for SSR and SNP datasets). Levels of expected and observed heterozygosity for the SSR dataset were only slightly greater than those estimated for the SNP dataset. Likewise, the estimated mean effective population size per sampled locality was only slightly greater and more variable for the SSR dataset than for the SNP dataset (13.71±5.59, 9.49±0.13, respectively), but in neither case was there salient evidence of a plausible source-sink dynamic, as judged by the similar effective sizes among populations. Neither were there salient differences in $F_{ST}$ estimates between datasets (0.151 and 0.140, respectively for SSR and SNP datasets; Fig. S1), with $F_{ST}$ estimates being within the range of $F_{ST}$ values of other broadly distributed agricultural weeds [$F_{ST}$: 0.14–0.38] (Bussell, 1999; Eschmann-Grupe, Neuffer, & Hurka, 2004; Müller-Schärer & Fischer, 2001). Congruently, no major differences in genetic estimates between the SSRe and SNPC estimates were found (Table S4). Further confirming the limited spatial structure in this species, spatial genetic clusters identified by the best TESS model (Fig. S2) explained less than 13% of the variance across datasets, and barely reduced the variance explained solely by geographic location when compared to a null model with no regions assigned (Tables 1, S5).

Despite these similarities between the SSR and SNP datasets, the underlying genetic structure was markedly different. Estimates of recent ancestry differed between SSR and SNP datasets. The analysis on the SSR dataset indicated that migration among localities is more widespread and hardly geographically constrained, with only four localities being primarily constituted of native individuals (Fig. 2a). Across localities, on average 73.65% of individuals...
were inferred to be 1st or 2nd generation immigrants (it is important to note, however, that such high migration rate surpasses the assumptions of the method, and hence they should be taken cautiously). On the other hand, the analysis of the SNP dataset showed that most populations have a much more limited number of recent immigrants, and that the relatively few inferred immigrants (on average 27.42% of individuals) did not come exclusively from geographically proximate localities (Fig. 2d). Accordingly, SSR and SNP pruned conditional genetic networks (Dyer & Nason, 2004) indicated remarkably different underlying patterns of genetic connectivity (structural congruence = 0.108; Fig. 2b,e). The SSR-based network was more interconnected (vertex connectivity: 5) than the SNP-based network (vertex connectivity: 0). Furthermore, strong admixture was recovered in the SSR dataset, whereas minimal admixture was identified in the SNP dataset (Fig. 2c,f). As before, these differences were consistent when analyzing the SSRc and SNPC datasets. The SNPC dataset was characterized by a smaller percentage of recent immigrants (28.25%) than the SSRc dataset (44.93%) (Fig. S3a,d), and the corresponding genetic networks were also clearly different from each other (structural congruence = 0.002)—with the SSR-based network being more connected (and vertex connectivity = 2) than the SNPC-based network (i.e., vertex connectivity = 0)—(Fig. S3b,e). Finally, as for the full data, a more admixed genetic composition of individuals was recovered in the SSRc dataset than in the SNPC dataset (Fig. S3c,f).

Landscape genetics

Unsurprisingly, given the distinct underlying genetic patterns between SSR and SNP datasets (see above), the optimization of landscape resistance layers resulted in different resistance optimization solutions for each dataset (Fig. S4). It is important to note, however, that a formal comparison is unwarranted as the associations recovered are statistical associations.
driven by the fit of the resistance parameterization to the data under the statistical model implemented (Martínez-Abraín, 2008). While these associations are expected to recapitulate real biological properties of the study system, they are constrained to the data at hand. Nonetheless, association patterns that are robust to the data used are expected to better reflect the actual impact that landscape features have on gene flow, independently of possible biases introduced by expert opinion. Therefore, we focus below on the common biological findings between marker types, while also denoting the most relevant differences. Such differences likely reflect not only the different environmental ranges covered by each dataset (Fig. 1), but most importantly, the particular population genetic structure underlying each dataset (Fig. 2).

In spite of the distinct underlying genetic structure, there were some landscape features that showed consistent inferred conductance to migration between datasets (Fig. S5). For example, a tendency towards greater landscape conductivity in relatively warm and precipitation-seasonal areas was observed in both datasets along with a steep decline in connectivity towards areas with the greatest temperatures and precipitation-seasonality values in the study area (Fig. S5). Likewise, cotton-dominated areas were recovered as substantially more permeable landscape features than areas dominated by soybean fields, evergreen forests, open shrublands, and grasslands (Fig. S5). Furthermore, both datasets pointed towards human-impacted landscapes playing an important role in shaping genetic connectivity in this species. While in both sets of MLPE.lmm models, null (geographic distance), natural (climate, elevation, and soils), and human-related landscapes (landcover and human population density) were identified as significant (hereafter, 0.01<p<=0.05 after correction for multiple testing) or marginally significant predictors (hereafter, p<=0.01 after correction for multiple testing) of genetic similarity between localities, the variable with the greatest association coefficient and lowest
AICc value in these models was in both cases a variable closely linked to human presence (landcover in the SSR dataset, and human population density in the SNP dataset; Table 2) (for comparison, Table S6 shows comparable analyses based on distance-based Redundancy Analysis—another commonly used algorithm in landscape genetics). However, when considering all variables together in a multivariate manner while accounting for geographic distance, human population density, local effective population size, and different aspects of climate were the only variables that remained as significant or marginally significant predictors of genetic differentiation across both SSR and SNP datasets (Table 2). Elevation and soil were identified as significant or marginally significant predictors only in the SNP dataset.

In summary, across datasets results indicated that human-population-density resistance was robustly associated with population differentiation, with highly populated areas identified as less conducive areas for migration (although the exact association varied between datasets; Fig. S5). Local effective population size was also a significant predictor when considering all other variables. It is important to note, however, that these multivariate regressions explained a variable proportion of the variance (MRDM $R^2$ for SSR and SNP dataset were 0.109 ($F_{1,29} = 3.654$, p-val. = 0.063) and 0.532 ($F_{1,6} = 1.932$, p-val. = 0.113), respectively). In addition to population density and effective population size, temperature was also recovered as a significant predictor of genetic dissimilarity across most analyses for the SSR dataset but not for the SNP dataset (Table 2).

**DISCUSSION**

The results suggest that broadly distributed populations of *I. purpurea* are partially genetically distinct (more so when analyzing the SNP dataset), although there is some indication of long-distance and putatively human-mediated migration between localities—as suggested by
the recovered association between human population density and genetic similarity. The levels of
differentiation observed and inferred long-distance migration strongly contrast with this species’
patchy distribution, which is tightly linked to isolated agricultural patches that are surrounded by
a complex matrix of natural and urbanized areas. Contrary to what has been seen in other
agricultural weeds (Menchari et al., 2007; Ye, Mu, Cao, & Ge, 2004), populations of *I. purpurea*
in relatively close geographic proximity do not form clusters of genetically similar individuals.
This finding, supported by our spatial population structure and landscape genetics analyses,
suggests that the local agricultural matrix does not seem to have an overarching impact on the
connectivity in this species at the landscape level—albeit it likely influences connectivity at the
small spatial scales. Instead, climate and human population density were robustly recovered as
predictors of genetic connectivity in this species across datasets and analyses. Of these landscape
variables, climate has a stronger effect, as judged by its greater MRDM coefficient. Of note,
temperature (summarized by climate PC1) was recovered as marginally significant only when
considering the SSR dataset, which is the only dataset that covers the northern portion of the
range, whereas precipitation seasonality (summarized by climate PC2) was recovered as
marginally significant only in the SNP dataset. Otherwise, population density was the only
variable across datasets with a marginally significant effect—even after accounting for multiple
tests. In addition, local effective population size was found to be a significant predictor only after
accounting for all landscape variables, suggesting a plausible superseding effect of genetic drift
driving differentiation across localities (Weckworth et al., 2013). Taken together, these results
highlight the significant interplay between human-driven and natural landscapes in structuring
populations of this species. The role that humans play in this system is likely mediated by their
impact on migration patterns themselves as well as by the reduction of population size of this
weed through pest control (Barker, Thompson, & Godley, 1984; Baucom & Mauricio, 2008).
The results also highlight how inferences about population structure and patterns of connectivity may be dataset-dependent, with marked differences becoming apparent only after careful dissection of roughly similar $F_{ST}$ and heterozygosity estimates across molecular markers. Such differences cannot be attributed to the more widespread samples of our SSR dataset, as our findings were robust to subsampling this dataset to match the available SNP samples (see Supporting Information). This represents a rather unexpected finding as both the overall population structure and the influence of landscape features on population connectivity should be inherent species properties and no marker specific realizations of common underlying biological processes (but see below). Next, we detail each of these novel findings and place them in the context of agricultural weed movement across the landscape, invasive species, and landscape genetics practice.

**Human impact**

Given that *I. purpurea* is a naturalized species in the United States that is found primarily associated with cultivated crops and horticultural gardens (Baucom & Mauricio, 2004; Defelice, 2001; Fang et al., 2013), the finding that human population density is a predictor of genetic similarity in this species is at first glance unsurprising. Yet, because habitat requirements for establishment and migration are not always the same, especially for organisms with distinct migration stages (e.g., pollen or seeds in plants) and dormant stages (Murphy & Lovett-Doust, 2004), this finding is not as straightforward as it seems. In particular, the fact that human population density is recovered as an informative predictor throughout the entire sampled distribution—even after accounting for climate and landcover variation, highlights the direct influence that humans most likely have on structuring the populations of this species and helps to
discern the factors involved in the spread of this noxious weed. In this sense, the results point towards humans not only as likely responsible for the introduction of this weed into the United States (Fang et al., 2013), but also as likely responsible for facilitating its current spatial connectivity and genetic structure, and hence its opportunities for thriving in the complex landscapes it inhabits. While further testing is required to formally test this hypothesis, especially considering the limitations of current landscape genetic approaches (see below), our findings suggest a multifaceted effect of human activities.

For one, it is theoretically possible that human population density primarily facilitates connectivity at local to intermediate scales, which encompasses agricultural fields in relatively close geographic proximity, suggesting that factors such as regional sharing of contaminated agricultural machines, regional trade between farmers, or regional distribution of contaminated crop seeds were at play (Benvenuti, 2007; Boyd & White, 2009; Dastgheib, 1989; Thill & Mallory-Smith, 1997). Yet, the limited spatial clustering at the regional scale and the lack of a significant effect of roads (results not shown) make this possibility unlikely. Instead, considering that i) the horticultural trade has been recognized as the main source of invasive introductions and spread in the United States (Lehan, Murphy, Thorburn, & Bradley, 2013), ii) that *I. purpurea* is an appreciated horticultural species (Fang et al., 2013), and that, given current agricultural practices, crop seed contamination is unlikely to be a major factor (Economic Research Service, 1998), it is probable that ornamentals’ trade between population centers may help explain both the long distance dispersal events recovered in both datasets and the overall population structure. Alternatively, the impact of human populations on the distribution and abundance of bumblebees (Jha, 2015; Martins, Goncalves, & Melo, 2013), which are *I. purpurea*’s predominant pollinators (Baucom & Mauricio, 2008; Ennos, 1981), could also be partially responsible for the landscape
genetic patterns recovered as changes in the pollinators community would have strong effects on gene flow (Jha & Kremen, 2013).

In reality a combination of all these factors may be involved. While further analyses are needed to elucidate the ultimate causes behind the recovered association between human population density and genetic dissimilarity in *I. purpurea*, our findings bring much needed insight to limit the spread of this noxious weed. Our findings are not only relevant to *I. purpurea* and to the evolution of herbicide resistance in this species (i.e., is herbicide resistance evolving independently across populations or is it being disseminated through human-aided migration?), but also have important implications for other weeds of agricultural concern as well as other human-exploiter species (Blair, 2001), such as other invasives. Specifically, in line with previous work (Auffret, Berg, & Cousins, 2014; Banks, Paini, Bayliss, & Hodda, 2015; Bataille et al., 2011), the results here point towards the need of better strategies to minimize the impact that humans have on the spread of species. In particular, our results further support that humans may not only facilitate the introduction of invasive species into non-colonized areas, but also contribute to the maintenance of gene flow among naturalized populations (Medley, Jenkins, & Hoffman, 2015), which may be critical in providing relevant genetic variants to respond to novel selective regimes as well as prevent inbreeding depression in these newly colonized areas (Edelaar & Bolnick, 2012; Kolbe et al., 2004).

**SSR- vs. SNP-based inferences**

The unique patterns observed for each marker offer the opportunity to explore the underlying causes for such differences and hence a more in-depth understanding of the plausible landscape influences on species’ genetic structure. For instance, an important consideration in any
landscape genetics study, including this one, is the spatial distribution of samples and spatial scale of environmental data (Wang & Bradburd, 2014), as it can strongly impact the associations recovered. It is thus theoretically possible that the particular geographic sampling of each dataset exclusively drives the differences in genetic structure recovered by the two markers. Yet, the robust differences that we report between the localities and regions common to both datasets (SSRc and SNPC subsampled datasets; see Supporting Information) renders this possibility highly unlikely and suggest that our results are at least moderately robust to sample distribution. Still, it is important to recognize that all our datasets contain sets of spatially clustered samples, partially in reflection of the also spatially cluster distribution of agricultural fields (Ramankutty, Evan, Monfreda, & Foley, 2008). Hence, it is in principle possible that our inferences on all 4 datasets might be strongly impacted by the lack of samples from intervening areas. However, our analyses do not show the pattern of genetic separation between geographic sample clusters that is expected under clustered sampling (Schwartz & McKelvey, 2009), suggesting that the patterns recovered are not simply a sampling artifact.

Instead, differences between SSR- and SNP-based patterns might be related to the different mutation rates underlying the two type of markers (Wang, 2010, but see Bohonak and Vandergast 2011). SSR mutation rates per generation per site ($\mu$) are typically estimated to be between $10^{-3}$ and $10^{-4}$ mutations per genome site per generation (Garza & Freimer, 1996), whereas SNP mutation rates are typically estimated to be on average orders of magnitudes slower, around $10^{-8}$-$10^{-10}$ (Morin, Luikart, & Wayne, 2004). All else being equal, the inferential power of population structure is tightly linked to the number of mutations (Hubisz, Falush, Stephens, & Pritchard, 2009; Turakulov & Easteal, 2003). As a consequence, unless there is widespread homoplasy it is expected to be more likely to recover signatures of population differentiation using the faster
evolving SSR loci. This is true even considering the total number of loci on each dataset (15 SSR loci vs. 8210 SNP loci) (Selkoe & Toonen, 2006). Yet, our results are in contrast with this theoretical expectation as we recovered weaker population structure using the faster evolving SSR loci than using SNP loci. It is still possible that the greater number of expected mutations for SSR loci, which increase the opportunities for homoplasy (Garza & Freimer, 1996), may explain the lower degree of population differentiation in this dataset. Nonetheless, the likelihood that widespread loci homoplasy in SSR allele size across populations has been maintained over the temporal window since the introduction of *I. purpurea* to the US seems small. Hence, this mutation-differential hypothesis is unlikely to be solely responsible for the differences observed. In fact, a large proportion of genetic variation in current populations might, depending on effective population size and generation time, precede the temporal window of many landscape genetics studies. In this regard, the greater number of SNP loci translates into a better genomic representation. Since different genomic regions reflect different coalescent histories (Nielsen & Slatkin, 2013), increased genomic coverage should better capture the range of processes conditioning genetic patterns of populations and thus the combined effect of historical demographic processes and current landscapes. It is then important to consider the relative contribution of both processes: i) input from new mutations and ii) sorting of standing genetic variation, rather than exclusively focus on mutation rates differences. Such sorting is expected to be specific to different genomic regions, which most likely contribute to the differences observed between our SSR and SNP datasets.

**Advancing landscape genetics practice**

Signals of population structure may arise from a wide range of evolutionary processes, including historical demographic events (He, Edwards, & Knowles, 2013), local adaptation
(Orsini, Vanoverbeke, Swillen, Mergeay, & De Meester, 2013), and reproductive strategies (e.g., selfing in mixed mating species such as *I. purpurea* can lead to a spurious identification of structure; Gao, Williamson, & Bustamante, 2007). Yet, landscape genetics approaches traditionally overlook these plausible confounding processes by working under the assumptions of an equilibrium between migration and genetic drift and an implicit predominance of recent landscape configurations over alternative explanations for the observed population structure (Dyer, 2015b; He et al., 2013; Marko & Hart, 2011). Thus, traditional landscape genetics analyses presumably present an incomplete picture of the evolutionary processes driving current patterns of genetic diversity (Wang & Bradburd, 2014). Nonetheless, these approaches undoubtedly offer a valuable hypothesis-generation framework about the possible role that environmental setting plays in structuring genetic diversity against which the effect of other demographic processes can be evaluated. Arguably, the integration of landscape genetics with historical demographic reconstruction is key for robust population genetics inference since disregarding the plausible effects of either current landscape processes or historical demographic changes would impair the ability to understand species’ complex responses to spatio-temporal environmental variation.

In this context, considering the likely complex demographic dynamics of this introduced agricultural weed, our results should be taken as a working hypothesis of the possible role of the interaction between natural and anthropogenic landscapes in structuring *I. purpurea* populations. Nonetheless, our analyses represent a step towards integrating traditional landscape genetics with modern population genetics inference by taking advantage of recent analytical developments and richer molecular datasets. On one hand, our analyses use novel methodological tools that i) surpass the need of arbitrary landscape resistance assignment that make inferences sensitive to
subjectivity of expert opinion (Dyer, 2015a), ii) account for the indirect genetic similarity of populations (Dyer & Nason, 2004), iii) use rigorous statistical inferences (Balkenhol et al., 2009; Peterman et al., 2014), and account for plausible confounding processes (i.e., local effective population size; Weckworth et al., 2013)—although more work on accounting for additional processes such as historical demographic changes is needed. On the other hand, in contrast to the common practice in the field of using a single analysis (commonly Mantel test; Guillot & Rouset, 2013) and one or a few loci (although a few notable exceptions exist; e.g. Perry et al., 2013), which prevents an assessment of common patterns across the genome (Bohonak & Vandergast, 2011), our inferences are derived from common findings among two rather different sets of molecular markers. In doing so, we provide not only statistically robust inferences, but also a better representation of the genome. Hence, our inferences are not only less sensitive to ascertainment bias (Brandström & Ellegren, 2008) and coalescent and mutational variance (Buschiazzo & Gemmell, 2006; Nielsen & Slatkin, 2013; Steiner, Putnam, Hoeck, & Ryder, 2013), but have also the ability to uncover differences in the underlying population dynamics. Such differences have strikingly important implications. For example, when evaluating plausible approaches to the threat of an invasive species such as I. purpurea, recommendations would be quite different depending on whether gene flow is believed to be relatively widespread (as inferred by the SSR dataset) or whether it is believed to be minimal (as inferred by the SNP dataset). In this example, it is clear that knowledge-based management would clearly benefit from recognizing the current uncertainty in regards to the exact population connectivity as opposed to automatically relying on a single-marker story.

Given recent advances in next generation sequencing, it seems straightforward to focus on landscape genomics instead of few loci. Hence, development of methods for explicitly integrating
inferences from multiple genome regions and marker types, as it is customary in population genetics, would be of great value. By incorporating multiple loci and coupling traditional landscape genetic tools with coalescent-based simulations to explicitly model landscape effects on genetic population structure, a robust hypothesis framework could be develop to simultaneously account for both current landscape processes and demographic history of species (Balkenhol & Landguth, 2011; Hoban, Bertorelle, & Gaggiotti, 2012). Advances in this area are already being developed with promising perspectives (Alvarado-Serrano & Knowles, 2014; Harris et al., 2016; He et al., 2013).

**Final remarks**

By offering a working hypothesis of the effect of current landscapes on genetic differentiation, traditional landscape genetics results serve the purpose of identifying relevant models for further testing (Baguette, Blanchet, Legrand, Stevens, & Turlure, 2013; Dyer, 2015a). Under this framework, our results pave the way for rigorous simulation-based assessments of the role of landscape features in promoting or deterring population differentiation in a noxious agricultural weed, and hence for successful knowledge-based invasive management. Specifically, we identify a probably major role of human-driven gene flow and long distance dispersal events in the demographic history of this species. If this weed was, as hypothesized, singly introduced through horticulture from a European bottlenecked population during the European colonization of North America (Fang et al., 2013), distinct ancestral structure (pre-dating US introduction) would be unlikely. Under this scenario, the rather distinct clustering of individual subpopulations recovered would likely reflect the connectivity driven by agricultural and horticultural activities and the complex natural/human landscape *I. purpurea* has experienced post-introduction to the US. Alternatively, human trade might have allowed for recurrent introduction events, which
effects would have probably been amplified by local agricultural and horticultural activities.

Regardless, these findings call for future model-based inference that explicitly considers the impact of human population density in conjunction with climate to further investigate the evolutionary drivers of population structure in this noxious weed.

ACKNOWLEDGEMENTS

We thank Adam Kuester for seed collecting and for contributing valuable data for this study. We also thank Ariana Wilson, Eva Fall, and Dan York for tissue collection. This research was funded by USDA NIFA grants 04180 and 07191 to R.S.B.

DATA ARCHIVING

All data generated is in the process of being archived in Dryad.

AUTHOR CONTRIBUTIONS

D.F.A.-S. and R.S.B conceived the study. D.F.A.-S. and M.V.E. generated and compiled the molecular and GIS data. D.F.A.-S. analyzed the data. D.F.A.-S., M.V.E., S.M.C., and R.S.B. wrote the manuscript. All authors read and approved the final submission.
LITERATURE CITED


Bussell, J. D. (1999). The distribution of random amplified polymorphic DNA (RAPD) diversity


polymorphisms within substructured populations of morning glory (*Ipomoea purpurea*). *The American Naturalist, 128*, 840–858.


of the Royal Society B, 276, 523–532.


### TABLE 1

Analysis of Molecular Variance (AMOVA) of SSR and SNP data. The contribution of spatial clusters (regions), localities, and individuals is shown. For comparison, results from an AMOVA analysis with no region category defined are presented in parentheses underneath.

<table>
<thead>
<tr>
<th>Effect</th>
<th>F-statistic</th>
<th>Variance explained</th>
<th>F-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SSR</td>
<td>SNP</td>
<td>SSR</td>
</tr>
<tr>
<td>Regions</td>
<td>F_{RT}</td>
<td>3.94%</td>
<td>8.51%</td>
<td>0.006</td>
</tr>
<tr>
<td>Localities</td>
<td>F_{SR}</td>
<td>9.05%</td>
<td>(11.06%)</td>
<td>6.10%</td>
</tr>
<tr>
<td>Individuals</td>
<td>F_{ST}</td>
<td>0.67%</td>
<td>(38.33%)</td>
<td>24.85%</td>
</tr>
<tr>
<td>(among)</td>
<td></td>
<td></td>
<td></td>
<td>0.146</td>
</tr>
<tr>
<td>Individuals</td>
<td>F_{IS}</td>
<td>86.33%</td>
<td>(50.61%)</td>
<td>60.54%</td>
</tr>
<tr>
<td>(within)</td>
<td></td>
<td></td>
<td></td>
<td>0.291</td>
</tr>
<tr>
<td>Total</td>
<td>F_{IT}</td>
<td>100%</td>
<td>(100%)</td>
<td>100%</td>
</tr>
</tbody>
</table>


Table 2. Summary of landscape genetics models. Model coefficients are reported followed by associated p-value (in parenthesis) and, for MLPE.lmm models, followed by AICc difference and ranking (in square brackets). Significant coefficients are in bold, marginally significant coefficients are marked with an asterisk.

<table>
<thead>
<tr>
<th>Feature</th>
<th>MLPE.lmm</th>
<th>MRDM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSR</td>
<td>SNP</td>
</tr>
<tr>
<td><strong>Intrinsic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic distance</td>
<td>0.187* (0.052)</td>
<td>0.780* (0.055)</td>
</tr>
<tr>
<td></td>
<td>+12.303[8]</td>
<td>+0.396[4]</td>
</tr>
<tr>
<td>Population size (Ne)</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

| **Natural environment variables** |          |      |      |      |
| Climate PC1          | 0.243 (0.034) | 0.967 (0.045) | 0.533* (0.064) | 1.782 (0.517) |
| Climate PC2          | 0.205 (0.048) | 0.814* (0.055) | -0.029 (0.978) | 21.443* (0.075) |
| Elevation            | 0.244 (0.034) | 0.840* (0.054) | -0.607 (0.480) | -31.789* (0.095) |
| Soil PC1             | 0.208 (0.039) | 1.044 (0.045) | 0.305 (0.510) | 15.789 (0.038) |
| Soil PC2             | 0.320 (0.018) | 0.941 (0.045) | 0.187 (0.703) | -6.838 (0.476) |
|                    | +8.139[2]   | +0.284[3] |       |       |

<p>| <strong>Human-impact variables</strong> |          |      |      |      |
| Crops                 | -0.226 (0.134) | 0.858* (0.054) | 0.154 (0.526) | 0.360 (0.840) |
| Landcover            | 0.582        | 1.358 | 0.340 | 3.887 |</p>
<table>
<thead>
<tr>
<th></th>
<th>(0.184)</th>
<th>(0.218)</th>
<th>(0.003)</th>
<th>(&lt;0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population density</td>
<td>0.912</td>
<td>-0.519*</td>
<td>+39.775</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.095)</td>
<td>(0.003)</td>
<td>[1]</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>[9]</td>
<td>[1]</td>
</tr>
<tr>
<td></td>
<td>-0.519*</td>
<td>-3.271</td>
<td>0.912</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.034)</td>
<td>(0.045)</td>
<td>(0.034)</td>
</tr>
<tr>
<td></td>
<td>-3.271</td>
<td>-</td>
<td>0.912</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>-</td>
<td>(0.045)</td>
<td>(0.034)</td>
</tr>
<tr>
<td></td>
<td>-0.519*</td>
<td>-6.519</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.095)</td>
<td>(0.037)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
**FIGURES**

Figure 1. Distribution of *Ipomoea purpurea*’s sampled localities. Sample sizes for both SSR (black bars) and SNP (white bars) datasets are indicated (locality numbers are given in squares). Elevation is provided as background.
Figure 2. Inferred population connectivity. The estimated origin of individuals for each sampled locality (i.e., sink) is depicted according to the locality they were inferred to have originated from (source) (a, d). The color of each cell in these plots depicts the proportion of individuals in the sink population that were estimated to be recent immigrants from each locality along the x-axis. Cells on the minor diagonal correspond to the proportion of native individuals. Pruned conditional genetic networks (b, e) and posterior estimates of admixture proportion identified by TESS analysis (c, f) are also displayed. The top row shows SSR-based results, the bottom shows the SNP-based results. Locality numbers follow Fig. 1. Localities shared between SSR and SNP datasets are denoted by unique colors in all panels (for a corresponding figure based exclusively on these shared localities, see Fig. S3).