Species complex diversification by host plant use in an herbivorous insect: The source of Puerto Rican cactus mealybug pest and implications for biological control

Daniel Poveda-Martínez[1], María Aguirre[1], Guillermo Logarzo[1], Stephen Hight[2], Serguei Triapitsyn[3], Hilda Diaz-Sotero[4], Marcelo Vitorino[5], and Esteban Hasson[6]

[1]Fundación para el Estudio de Especies Invasivas
[2]U.S. Department of Agriculture-ARS
[3]University of California
[4]Caribbean Advisor to the APHIS Administrator, USDA
[5]Universidade Regional de Blumenau
[6]Instituto de Ecología, Genética y Evolución de Buenos Aires

May 6, 2020

Abstract

Cryptic taxa have often been observed in the form of host-associated species that diverged as the result of adaptation to alternate host plants. Untangling cryptic diversity in species complexes that encompass invasive species is a mandatory task for pest management. Moreover, investigating the evolutionary history of a species complex may help to understand the drivers of their diversification. The mealybug Hypogeococcus pungens was believed to be a polyphagous species from South America and has been reported as a pest devastating native cacti in Puerto Rico, also threatening cactus diversity in the Caribbean and North America. There is neither certainty about the identity of the pest, nor the source population from South America. Recent studies pointed to substantial genetic differentiation among local populations, suggesting that H. pungens is a species complex. In this study, we used a combination of genome-wide SNPs and mtDNA variation to investigate species diversity within H. pungens sensu lato to establish host plant ranges of each one of the putative members of the complex, to evaluate whether the pattern of host plant association drove diversification in the species complex, and to determine the source population of the Puerto Rican cactus pest. Our results suggested that H. pungens comprises at least five different species, each one strongly associated with specific host plants. We also established that the Puerto Rican cactus pest derives from southeastern Brazilian mealybugs. This is an important achievement because it will help to design reliable strategies for biological control using natural enemies of the pest from its native range.

1. INTRODUCTION

Herbivorous insects are often involved in close interactions with their hosts since they are a food source and provide mating and oviposition sites (Schoonhoven, Van Loon, & Dicke, 2005). Such intimacy often entails the evolution of adaptations that allow insects to cope with specific features of their host plants. Hence, host plant shifts may affect the evolution of features associated with feeding location, oviposition and development on the host (Schoonhoven, Van Loon, & Dicke, 2005; Orsucci, et al., 2018). Adaptation to new hosts may cause, either as a direct consequence or as a byproduct, the evolution of sexual isolation, highlighting the role of host plant shifts in speciation (Funk et al., 2006; Nosil, 2012).

Phytophagous insects with presumptively wide host ranges, i.e. generalists, pose a concern from an applied view when dealing with insect pests. Many authors have suggested that polyphagous species are formed by

locally adapted specialized populations or cryptic specialist species (e.g., Loxdale & Harvey 2016; Forbes et al., 2017). A species complex made up of cryptic specialists pose an additional complication in defining host range. The distinction between true polyphagous and cryptic specialist species is crucial for the design of biological control programs. In addition, cryptic species complexes offer the opportunity to investigate the role of host plants in the diversification of herbivorous insects since such complexes often include species of recent origin (Winter, Friedman, Astrin, Gottsberger, & Letsch, 2017; Malka et al., 2018; Bakovic et al., 2019).

Distinguishing between intra and interspecific genetic variation is particularly relevant in cases of suspected cryptic species, especially when members of a guild of cryptic species are involved in invasions of new areas. Invasive insects may cause global problems upon spread to new territories, not only to agriculture, but also to biological diversity since invasive species are among the main causes of biodiversity loss (Newbold et al., 2015). The proper identification of invasive species is a necessary task to design successful biological control strategies to prevent or reduce harmful effects of invaders.

With the advent of molecular data, an increasing number of studies showed that some putative polyphagous insects are, actually, species complexes embracing several deeply diverged species, each one specialized to different host plant use (Egan, Nosil, & Funk, 2008; Nosil et al., 2012; Powell, Forbes, Hood, & Feder, 2014; but see Vidal et al., 2019). So far most population surveys have been based on the evaluation of a single genetic marker, the so-called DNA-barcoding gene encoding cytochrome oxidase subunit I (mtDNA) (Dinsdale, Cook, Riginos, Buckley, & De Barro, 2010; Stouthamer et al., 2017). However, it is well known that in many cases this marker provides information limited to the mitochondrial lineage, potentially resulting in misidentification of species boundaries due either to incomplete lineage sorting or introgressive hybridization (Eyer, Seltzer, Reiner-Brodetzki, & Hefetz, 2017; Després, 2019). Recently, population genetic studies benefited from availability of new methodologies based on high-throughput sequencing of genomic libraries containing a reduced-representation of nuclear genomes, known as genotype by sequencing, such as RADseq (Baird et al., 2008; Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). These methodologies provide large multi-locus datasets that can be used to evaluate cryptic diversity in species complexes (e.g., in Elfekih et al., 2018), and to trace the origin of invading pests (Anderson, Tay, McGaughran, Gordon, & Walsh, 2016; Ryan et al., 2019).

The cactus mealybug pest invading Puerto Rico and the adjacent smaller islands, represents a threat for cactus diversity in the Caribbean, Central and North America. The pest was initially reported as *Hypogeococcus pungens* Granara de Willink (Hemiptera: Pseudococcidae), under the common name Harrisia cactus mealybug (HCM), the successful biological control agent released against an invasive cactus in Australia and South Africa (McFadyen & Tomley 1978, 1981; Paterson, Hoffmann, Klein, Mathenge, Neser, & Zimmermann, 2011). Currently, *H. pungens sensu lato* is considered a species complex, and in its native range (South America) it was reported to feed on species of Amaranthaceae, Portulacaceae and Cactaceae (Ben-Dov, 1994; Claps & Haro, 2001; Zimmermann, Pérez Sandi Cuen, Mandujano, & Golubov, 2010). Recent studies based on molecular data and assessing reproductive compatibility (Poveda-Martínez et al., 2019) and performance on different hosts plants (Aguirre et al., 2016) suggests that populations collected in Argentina, initially identified as *H. pungens sensu lato* , were, actually, a species complex comprising at least two species; one mealybug species feeding on Amaranthaceae (*H. pungens sensu stricto*), and the other an undescribed species feeding on cactus. Interestingly, the Puerto Rican cactus pest appears closer to *H. pungens sensu stricto* in phylogenetic trees suggesting that the pest shares a recent ancestor with the latter rather than with the sympatric, cactus feeding new species from Argentina. However, none of the mitochondrial haplotypes found in Argentina matched the single haplotype detected in the Puerto Rican mealybugs feeding on Cactaceae, suggesting that the source population of the pest was not Argentina (Poveda-Martínez et al., 2019).

In the present study, we extended the sampling effort to a large geographic area comprising both the native (Argentina, Paraguay and Brazil) and non-native (Puerto Rico and southern United States) ranges by collecting mealybugs on host plants recognized as part of the diet of *H. pungens sensu lato* . We used a combination of genome-wide SNPs and mtDNA variation to investigate: i) genetic diversity within *H. pungens*

species complex, ii) host plant ranges of each one of the putative members of the complex; iii) whether host plant shifts drove the diversification in the species complex, and iv) the source population of the Puerto Rican cactus pest. Based on these results, we will be able to search for and select biological control strategies using natural enemies, either those that co-evolved with the pest (classical biological control) or those that did not co-evolve but attack closely related species to the pest (new association biological control).

2. METHODS

2.1 Sample collections and DNA extraction

Samples of the mealybugs were collected in the native range of *H. pungens sensu lato* in South America (n= 80) and in the recently invaded areas of Puerto Rico and the continental United States (n= 93). Mealybugs were collected on the host plants reported as part of the diet of *H. pungens sensu lato* (Cactaceae, n= 89; Amaranthaceae, n= 69; and Portulacaceae, n= 25). In the native range, mealybugs were collected in northern and northwestern Argentina (n= 21 on Cactaceae and n= 18 on Amaranthaceae), along the Atlantic coast of Brazil (n= 23 on Cactaceae; n= 11 on Amaranthaceae and n=4 on Portulacaceae), and in western Paraguay (n= 3 on Cactaceae). We also included samples collected in the non-native range in Puerto Rico (n= 42 on Cactaceae; 24 on Amaranthaceae and n= 11 on Portulacaceae), southeastern United States (n= 11 on Amaranthaceae), and in southwestern United States (n= 5 on Cactaceae). Additionally, we included samples from southeastern Australia (n= 5 on Cactaceae), where *H. pungens sensu lato* from Argentina was introduced as a biological control agent against cactus weeds in the 1970s (McFadyen, 2012). All individuals were preserved in 100% ethanol and stored in a freezer until DNA extraction. Information concerning sampling localities, geographic coordinates, and host plants species are presented in Figure 1 and Table S1.

Genomic DNA was extracted using entire bodies of adult female mealybugs using Qiagen DNeasy Blood & Tissue Kit according to manufacturer's instructions (Valencia, CA, USA), adding 2 µL of RNAse A after the lysis step. DNA was quantified using Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA), and quality was assessed in a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies Inc., Wilmington, DE, USA).

2.2 NextRAD sequencing

Genomic DNA was converted into nextRAD genotyping-by-sequencing libraries (SNPsaurus, LLC, Eugene, OR, USA) as in Russello, Waterhouse, Etter, and Johnson, (2015). Genomic DNAs were first fragmented with Nextera DNA Flex reagent (Illumina, Inc., San Diego, CA, USA), which also ligates short adapter sequences to the ends of the fragments. The Nextera reaction was scaled at a 1/5th reaction, and 5 ng of genomic DNA was used due to the presence of inhibitors in the samples. Fragmented DNAs were then PCR amplified for 26 cycles at 73^{o}C, with one of the primers matching the adapter and extending 9 nucleotides into the genomic DNA with the selective sequence GTGTAGAGC. Thus, only fragments starting with a sequence that matched the selective sequence of the primer were efficiently amplified. NextRAD libraries were sequenced in a 150 bp single-end reads mode in a lane of a HiSeq 4000 (SNPSaurus at University of Oregon). Raw sequence data are available at the National Centre for Biotechnology Information (NCBI), under the BioProject accession number PRJNA593002.

2.3 SNP calling and genomic data filtering

The genotyping analysis used custom scripts (SNPsaurus, LLC) that trimmed the reads using bbduk (BBMap, http://sourceforge.net/projects/bbmap/) (Bushnell, 2017) to remove nextera adaptors and low-quality sequences. A reference draft genome was built using DNA from a single male mealybug (Table S1). By means of this procedure we were able to obtain a clean reference since males do not feed as adults, allowing us to identify reads of bacteria, parasitoids or food scraps present in females' sequences. To build the reference genome, 150 bp paired-end reads were sequenced in a lane of a HiSeq 4000 (SNPSaurus at University of Oregon). Illumina paired-end sequences were then trimmed for Nextera adapters using bbduk (BBMap, sourceforge.net/projects/bbmap) (Bushnell, 2017). The assembly was done using abyss-pe (Jackman et al.,

2017). Assembled contigs shorter than 250 bp were removed and then aligned using blastn to the NCBI nt database. Blast hits to bacterial species were removed. Reference draft genome is available as FuEDEI_-HPun_1.1.fa at the NCBI under accession number: JAAOIU000000000. Cleaned reads were then mapped to the reference mealybug draft genome with an alignment identity threshold of 0.95 using bbmap (BBMap tools). Genotype calling was done by using callvariants (BBMap tools). The Variant Calling File (VCF) generated was filtered to remove individuals with more than 10% missing data, sites with more than 20% missing genotypes and loci with minor allele frequency (MAF) lower than 0.01 using VCFtools v1.15 (Danecek et al., 2011). We randomly selected one SNP per contig to minimize linkage disequilibrium (LD) and to ensure the independence of the SNPs employed in the next analyses. We also excluded from our dataset SNPs with more than two allele variants and indels. We used Bayescan to test for outlier SNPs. Fit to Hardy-Weinberg expectations (HWE) of variant frequencies for each locus within populations was tested using the exact test implemented in dDocent (Puritz, Hollenbeck, & Gold, 2014).

2.4 Mitochondrial gene amplification and sequencing

A fragment of the mitochondrial gene encoding for the Cytochrome oxidase subunit I *(COI*) was amplified using primers C1-J-2183 Jerry (Simon et al., 1994) and C1-N-2568 BEN3R (Brady, Gadau, & Ward, 2000). PCR reactions were performed following Poveda-Martínez et al. (2019). PCR products were checked in 1% agarose gels and purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany). Finally, both strands of the amplicons were sequenced using Sanger methodology at Macrogen Inc. (Seoul, South Korea). Additional *COI* haplotypes detected in Poveda-Martínez et al., (2019) were retrieved from GenBank (MN013440 - MN013554). All sequences were employed in phylogenetic and population genetic analyses. New mtDNA haplotypes were deposited in GenBank under access numbers: MT138921 - MT138931.

2.5 Population genomics, clustering and phylogeographic analyses

Based on SNPs datasets, observed heterozygosity (Ho) and allele richness (Ar) were calculated with R package *hierfstat* (Goudet, 2005). Expected heterozygosity (He), global differentiation estimates (Gst) and pairwise Fst estimates between populations were calculated using*mmod* and *Adegenet* (Jombart, 2008) packages available in R (www.r-project.org). Populations were defined by sampling location and host plant collection site; however, samples from Argentina, Australia and Paraguay were pooled due to low sample size after SNPs filtering, and also because Australian mealybugs have been shown to be part of the same clade with mealybugs from Argentina feeding on Cactaceae (Poveda-Martínez et al., 2019). To assess population structure, we employed two approaches. First, we used sNMF (Frichot, Mathieu, Trouillon, Bouchard, & François, 2014) as implemented in R using the LEA package (Frichot & François, 2015) to assign samples to genetic clusters. sNMF estimates individual ancestry coefficients utilizing the same likelihood model underlying structure (Pritchard, Stephens, & Donnelly, 2000) and admixture (Alexander, Novembre, & Lange, 2009), but uses nonnegative matrix factorization and least squares optimization to accommodate genome scale datasets. Ancestry coefficients were calculated by sNMF for K values ranging from 1 to 10, with 100 repetitions for each value of K , and the optimal K value was assessed using the cross-entropy criterion based on the prediction of masked genotypes to evaluate the error of ancestry estimation. The second approach used to assess population structure was a discriminant analysis of principal components DAPC (Jombart, Devillard, & Balloux, 2010) as implemented in R in*adegenet* package, using K means clustering to identify the optimal number of clusters for K values ranging from 1 to 10, and estimating individual admixture coefficients. We then calculated Bayesian information criterion scores (BIC) to identify the K value with the best fit. Additional pairwise Fst estimates between groups identified in prior clustering analyses were performed using the SNPs dataset (142 individuals, 1,707 SNPs).

Mitochondrial DNA sequences were aligned using Clustal W implemented in MEGA v7 (Kumar, Stecher, & Tamura, 2016) and used to build a haplotype Network using Median Joining algorithm (Bandelt, Forster, & Röhl, 1999) implemented in PopArt v1.7.1 (Leigh & Bryant, 2015). Additionally, we calculated nucleotide diversity (π), haplotype diversity (Hd) and number of haplotypes (H). Pairwise Fst estimates between populations (as defined earlier) were calculated using DNAsp v6 (Rozas et al., 2017).

2.6 Influence of host plants and geographic distance on genetic divergence

We analyzed the influence of geographic distance and host plants on genetic divergence among *H. pungens* complex populations from the native range. Non-native populations in these analyses were excluded because samples collected in these locations reached the areas by human mediated dispersal and would distort the analyses. We used a combination of correlation tests and multiple regression analyses of distance matrices to analyze the importance of geographic distance and host plant use on genetic divergence.

In these analyses, we used pairwise Fst linearized values at population level based either on nuclear SNPs or mtDNA datasets to construct the matrix of genetic distances. To construct the geographic distance matrix, we used the coordinates (latitude and longitude) of each population (Table S1). Host plant distance matrix was built considering host plant families where mealybugs were collected: Cactaceae, Portulacaceae, and Amaranthaceae. Thus, we constructed a binary matrix with three categories (the host plant families) considering patterns of host plant use.

To analyze the influence of geographic distance and host plant use on genetic divergence we employed two Mantel tests (Mantel, 1967), first considering geographic and genetic matrices and second host plant and genetic distance matrices. We then conducted a partial Mantel test (Smouse, Jeffrey, Long, Robert, & Sokal, 1986) to evaluate the influence of host plants on genetic divergence, using pairwise Fst and host plant distance matrix, while controlling for geographic distance (Legendre & Fortin, 2010). For all correlation analyses we used the R package *vegan* v2.4.4 (Oksanen et al., 2007) and 9,999 permutations to assess statistical significance. We further tested the influence of geographic distance and host plant use on genetic divergence by performing matrix regressions using the function MRM from the package *ecodist* v2.0.1 (Goslee & Urban, 2007), and ran 9,999 permutation to assess the significance of regression coefficients.

2.7 Species tree reconstruction and species delimitation analyses

For SNPs and mtDNA datasets, maximum likelihood (ML) and bayesian inference (BI) approaches were performed. ML for both datasets was implemented in IQ-TREE v1.6.10 (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) using CIPRES (Miller, Pfeiffer, & Schwartz, 2010) with 10,000 ultrafast bootstrap (UFboot) iterations. ModelFinder estimated GTR as the best-fitted model according to Akaike Information Criterion (AIC) for SNPs dataset and GTR+Γ+I for mtDNA dataset. BI reconstruction was assessed using MrBayes v3.2.6 (Ronquist et al., 2012), employing the entire matrix of 1,707 SNPs and the mtDNA haplotypes following the same fit-model implemented in ML analyses. For each dataset, two independent runs were performed with four Markov Chain Monte Carlo (MCMC) for each run for 50 million generations, sampling every 1,000 generations. The first 25% of the runs were removed as burn-in, and stability and sufficient mixing of parameters (ESS>1,000 for SNPs and ESS>200, for mtDNA) were checked using Tracer v1.6 (Rambaut, Suchard, Xie, & Drummond, 2014). UFboot values and Posterior Probabilities (PP) resulted from ML and BI, respectively, were plotted on each node of the reconstructed tree of each dataset. Additionally, based on SNPs dataset, a species tree was estimated using the coalescent-based methods, SNAPP (Bryant, Bouckaert, Felsenstein, Rosenberg, & RoyChoudhury, 2012) implemented in Beast v2.5.2 (Bouckaert et al., 2014). In this case, we could only include two individuals, selecting individuals with the lowest numbers of missing data, from each one of the groups defined in clustering analyses (see the previous section) as a result of the high computational demand of the method. Due to the reduction of the number of individuals, some sites could become monomorphic either because of the reference or the alternative allele, and some sites might also have had no data left for one or more of the sampled localities. Therefore, we decided to filter out monomorphic sites and sites with no data for one or more populations with bcftools v1.7 (Li et al., 2009), resulting in a new matrix of 1,679 SNPs. We used the default model parameters in SNAPP for U and V equal to one and we ran the analysis for 10,000,000 MCMC generations, sampling every 1,000 generations. Stationarity and convergence of data was checked using Tracer v1.6 (Rambaut et al., 2014). The complete set of trees were visualized in Densitree v2.2.5 (Bouckaert & Heled, 2014), removing the first 10% of the trees as burn-in. Finally, we generated a maximum clade credibility tree using TreeAnnotator v1.7.5 (Drummond, Suchard, Xie, & Rambaut, 2012) to access the posterior probabilities of the resulted topology.

Bayes Factor Species Delimitation (BFD*) (Leaché, Fujita, Minin, & Bouckaert, 2014) analysis based on the SNPs dataset, was performed using the SNAPP and *Path Sampler* packages included in Beast v2.5.2 (Bouckaert et al., 2014). We estimated and compared the marginal likelihood estimates (MLE) for two models of species limits. Model 1 considered *H. pungens* as a complex of the following five species according to the results obtained with sNMF, DAPC, and the estimated best phylogenetic tree: i) *H. pungen* s *sensu stricto*(*Hypogeococcus* populations collected on Amaranthaceae in Argentina); ii) mealybugs feeding on Cactaceae from Argentina-Paraguay-Australia; iii) mealybugs feeding on Cactaceae from southeastern Brazil and Puerto Rico; iv) mealybugs feeding on Amaranthaceae and Portulacaceae from northeastern Brazil, Puerto Rico and the United States; and v) mealybugs feeding on Amaranthaceae from southeastern Brazil. Model 2 considered *H. pungens* as a complex comprising four species according to the major clades found with mtDNA: the same groups as in Model 1 with the difference that *H. pungen* s*sensu stricto* and mealybugs feeding on Amaranthaceae and Portulacaceae from northeastern Brazil, Puerto Rico and the United States appear as only one group. The MLE for each model was calculated using the same mutation rate as above, considering 12 steps with an MCMC length of 200,000 generations, a pre-burning of 20,000 and discarding the first 10% as burning. The MLEs for each model were used to calculate the Bayes factor (BF) test statistics $2MLE(model1) - MLE(model2)$ (Kass & Raftery, 1995).

Finally, two single locus methods of species delimitation were run based on mtDNA, the Generalized Mixed Yule Coalescent (GYMC) (Pons et al., 2006) and Bayesian Poisson tree processes (bPTP) (Zhang, Kapli, Pavlidis, & Stamatakis, 2013). For the GYMC analysis we first calculated an ultrametric tree in Beast v2.5.2 (Bouckaert et al., 2014). We used GTR+Γ+I model with an uncorrelated relaxed clock and constant size tree prior. MCMC was set to 10,000,000 generations, sampling every $1{,}000^{\text{th}}$ generation. The maximum credibility tree was generated in TreeAnnotator v1.7.5 (Drummond et al., 2012). The GYMC analysis was performed with the *SPLITS* package (Ezard, Fujisawa, & Barraclough, 2009) in R, using the ultrametric tree inferred by Beast. The bPTP analyses were performed in the bPTP server (http://species.h-its.org/) using the ultrametric tree in nexus format, with default values and using 200,000 MCMC generations.

3. RESULTS

3.1 NextRAD and mitochondrial data

The nextRAD dataset consisted, on average, of 1.6 million single-end reads per individual, with an average read length of 123 bp. The final assembled genome contained 250,856 fragments with a total length of 240.6 Mb. Considering a set consisting of 173 samples, we identified 14,446 loci. However, we had to remove 32 individuals from further analyses because of high numbers of missing data (<10% missing genotypes). From the remaining 141 individuals, we recovered a total of 5,383 loci. After filtering loci that were present in at least 90% of the individuals, we ended up with a final dataset containing 1,707 SNPs (average coverage 78X) used for further analyses. Numbers of loci removed in each filtering step are detailed in Table S2. Since the 10% of missing data criterium used for filtering the SNPs dataset is the strictest threshold, we evaluated other filtering thresholds that yielded essentially the same results in population genetic structure and species delimitation analyses.

The mtDNA dataset consisted of 233 sequences, 143 collected in the native range and 90 in the non-native range. Additionally, we included *COI* sequences of five individuals from California, United States on Cactaceae, but SNPs data at this location could not be generated. The length of the amplified *COI* fragment was 413 bp, with 371 conserved (89.83%) and 42 (10.16%) variable sites, distributed in 27 haplotypes. The mtDNA sequences are available in GenBank under accession numbers MT138921 - MT138931.

3.2 Genomic diversity and differentiation in the *H. pungens* species complex

Both SNPs and mtDNA datasets revealed intermediate levels of genetic diversity but great levels of differentiation among populations. Overall nuclear and mitochondrial genetic diversity were high across the native range and low in the non-native range (Table 1). Based on SNPs data, total mean expected heterozygosity (He= 0.061) was three times greater than observed heterozygosity (mean Ho= 0.029), probably as a consequence of differentiation among populations (see below). Expected heterozygosity varied greatly among

populations ranging from 0.267 in the native range (mean He= 0.102) to as low as 0.003 in the non-native area (mean He= 0.061). Allele richness was 20% higher in the native range (mean Ar = 1.240) than in recently invaded areas (mean Ar = 1.020), and no private alleles were detected.

The analysis of pairwise Fst values estimated for the SNPs dataset, considering the *a priori*defined populations, sampling area, and host plant family, showed great global genetic differentiation (Gst= 0.908). In the native range, Fst estimates ranged from 0.602 for the comparison between BRC (Brazilian mealybugs feeding on Cactaceae) and BRA (Brazilian mealybugs feeding on Amaranthaceae and Portulacaceae) to 0.827 in the comparison between ARA (Argentina mealybugs feeding on Amaranthaceae) and ARC (Argentina, Paraguay and Australia mealybugs feeding on Cactaceae). Genetic differentiation (Fst) in the non-native area varied from values as low as 0.078 between PRA (Puerto Rican mealybugs feeding on Amaranthaceae or Portulacaceae) and USA (United States mealybugs feeding on Amaranthaceae) to as high as 0.969 between PRA and PRC (the Puerto Rican cactus pest) (Table S3).

Clustering analyses also revealed strong population structuring in the*H. pungens* species complex, not necessarily associated with site location. Indeed, sNMF and DAPC analyses suggested that the *H. pungens* species complex comprises five ancestral populations / clusters (K) based on cross-entropy and BIC values (Figure 2A and B, respectively). The first cluster corresponds to what was previously called *H. pungens sensu stricto* (Ar-A) (Poveda-Martínez et al., 2019) that included Argentina (Ar) mealybugs feeding on Amaranthaceae (A). The second cluster included Cactaceae (C) feeding mealybugs native from Argentina, Paraguay (Pa), and mealybugs collected in Australia (Au) (ArPaAu-C). The third cluster included populations of Cactaceae feeding mealybugs from southern Brazil (Br) and the Puerto Rican (PR) cactus pest (BrPR-C), whereas the fourth cluster encompassed mealybugs feeding on Amaranthaceae and/or Portulacaceae (AP) species from northeastern Brazil along with populations from Puerto Rico and the United States (US) (BrPRUS-AP). The last cluster included mealybugs collected on Amaranthaceae in southeastern Brazil (Br-A). Admixture coefficients suggested no admixture between clusters and strong population genetic structure. Pairwise Fst values considering the five clusters described above revealed strong genetic differentiation, Fst estimates ranged from 0.567 in the comparison between Ar-A and BrPRUS-AP to 0.945 in the comparison between BrPRUS-AP and Br-A (Table S4).

Concerning the mtDNA dataset (Table 1), a total of 27 haplotypes were found in native and non-native ranges. All haplotypes were present in the native range, 18 haplotypes were detected in mealybug samples collected on Amaranthaceae and/or Portulacaceae hosts and 9 on Cactaceae hosts (mean Hd= 0.744). As expected, the number of haplotypes was substantially lower in non-native sites, where only three haplotypes were found (mean Hd= 0.174), two in samples collected on Amaranthaceae and/or Portulacaceae hosts and only one on Cactaceae hosts.

Four haplogroups, separated by at least two substitutions, can be recognized based on mtDNA data, each one characterized by the type of host plant (Amaranthaceae and/or Portulacaceae, or Cactaceae) and geographic origin (Figure 2C). The first haplogroup (H1), which contained the most frequent haplotype (haplotype MT138927) and has a star-like shape, is shared by Amaranthaceae-feeding mealybugs collected in Argentina, and specimens from Brazil, Puerto Rico and United States (Florida) feeding on Amaranthaceae and/or Portulacaceae. The second haplogroup (H2), which includes two haplotypes, both from mealybugs using Cactaceae as host plants, one (haplotype MT138924) is shared by the mealybugs from southeastern Brazil, Puerto Rico and United States (California), and the other (haplotype MT138923) only found in southeastern Brazil. The remaining two haplogroups are highly differentiated from H1 and H2. H3 included Cactaceae-feeding mealybugs collected in Argentina, Australia and Paraguay, and H4 Amaranthaceae-feeding mealybugs from southeastern Brazil.

3.3 Influence of host plant use and geographic distance in genetic variation

Analyses of the influence of host plants and geographic distance on genetic variation in the *H. pungens* species complex showed that host plant use was a better predictor of genetic distance than geographic distance (Table 2). Mantel correlation tests between host plant use and genetic distance matrices and between geographic

and genetic distance matrices using the nuclear SNPs dataset were significant ($r = 0.5254$; $p = 0.0017$ and r= 0.4;p= 0.0128, respectively). In addition, the partial Mantel test indicated that the correlation between host plant distance and genetic distance matrices remained significant when controlling for geographical distance ($r = 0.6133$; $p = 0.0001$). Likewise, the results of regressing the matrix of genetic distances on both host plant and geographic distances showed that both factors significantly affected genetic differentiation among sampling localities for nuclear SNPs (host plant coefficient= 0.28497, p= 0.0001; geographic coefficient = 0.00709, p= 0.0001). Host plant and geographic distance together explained 47.6% of genetic variation. In contrast, correlation and multiple regression tests based on mtDNA distance matrix on geographic and host plant distances yielded non-significant results (Table 2).

3.4 Species tree reconstruction and species delimitation analysis

Both ML and BI phylogenetic analyses and the species tree inferred by SNAPP, all of them based on the SNPs dataset, revealed five major well supported clades (Figure 3A and C). The first and second clades, both feeding on Amaranthaceae and/or Portulacaceae, appear closely related, the first encompassing *H. pungens sensu stricto* (clade Ar-A) and, the second, mealybugs from northeastern Brazil, Puerto Rico and United States (Florida) (clade BrPRUS-AP), respectively. The third and fourth clades comprised cactus feeding mealybugs: the third from Argentina, Paraguay and Australia (clade ArPaAu-C) and the fourth from southern Brazil and Puerto Rico (clade BrPR-C). The fifth clade was an independent branch formed by Amaranthaceae feeding mealybugs collected in southeastern Brazil (clade Br-A) (Figure 3A and C).

The phylogenetic tree obtained with the mtDNA dataset showed contrasting results with respect to nuclear SNPs (Figure 3B). Four major clades were observed in ML and BI trees, one encompassing *H. pungens sensu stricto* with populations from northeastern Brazil, Puerto Rico and the United States (Florida) feeding on Amaranthaceae and/or Portulacaceae (clade ArBrPRUS-AP). The clade including cactus feeding mealybugs from southeast Brazil and Puerto Rico (clade BrPR-C) appears as the sister clade of ArBrPRUS-AP. The third clade included populations from Argentina, Paraguay and Australia feeding on Cactaceae (clade ArPaAu-C) that appeared close to Amaranthaceae feeding mealybugs from southeastern Brazil (clade Br-A) (Figure 3B).

Species delimitation analysis based on the BFD method using SNPs data supported a model in which *H. pungens* was a complex of five species (Model 1, Table S5), supporting the picture depicted by the phylogenetic tree and species tree produced with SNPs data (Figure 3A and C). This means that the clade which included *H. pungens sensu stricto* (Ar-A) and the clade comprising mealybugs from northeastern Brazil, Puerto Rico and the United States (Florida) feeding on Amaranthaceae and/or Portulacaceae (BrPRUS-AP), as well as the clade of Amaranthaceae feeders from southeastern Brazil (Br-A) may be considered as three separate species. Moreover, the latter appeared as the sister clade of the two clades of cactus feeders, one from Argentina, Paraguay and Australia (clade ArPaAu-C) and the other from southeastern Brazil and Puerto Rico (clade BrPR-C).

GYMC and bPTP, the single locus species delimitation methods based on mtDNA, recovered six (GYMC) and 10 (bTPT) groups (Figure S1). In both cases, GYMC and bPTP delimited four of the five groups identified with SNPs data. These single locus methods did not consider the fifth group that consisted of *H. pungens sensu stricto* and the population from northeastern Brazil, Puerto Rico and the United States (Florida) on Amaranthaceae and/or Portulacaceae. Both methods delimited three (GYMC) and four (bPTP) additional groups in the clade conformed by mealybugs from Argentina, Paraguay, and Australia feeding on Cactaceae. Two additional splits were considered by bPTP, the first in *H. pungens sensu stricto* and the second one in mealybugs from southeastern Brazil on Amaranthaceae (Figure S1).

4. DISCUSSION

Our study confirmed that *H. pungens* , commonly called the Harrisia cactus mealybug (HCM), is not a single polyphagous species, but a species complex consisting of at least five species. Each member of the complex was associated with particular hosts, indicating a high degree of specificity at the host plant family level. Our genomic survey allowed the identification of the source population from which the Puerto Rico

cactus pest was derived. This is an important achievement since it will help to design reliable strategies for classical biological control using natural enemies, such as specialized parasitoids, associated with the pest in its native range.

4.1 *Hypogeococcus pungens* species complex and host plant use

The specimens used to describe *H. pungens* were originally collected on an Amaranthaceae host (Granara de Willink, 1981). Further collections extended its host range to the Cactaceae, Apocynaceae and Portulacaceae families (Ben-Dov, 1994; Claps & de Haro, 2001; Zimmermann, Pérez Sandi Cuen, Mandujano, & Golubov, 2010), suggesting that *H. pungens* was a polyphagous species. However, our population genomics approach pointed to a clear separation between the mealybugs associated with Amaranthaceae or Cactaceae, and others that indiscriminately use Amaranthaceae and Portulacaceae species as hosts. These results confirmed that *H. pungens* was not a polyphagous species, but a species complex composed of cryptic species associated with different host plants as feeding resources. Population genomics, clustering and species delimitation analyses based on genome wide SNPs and the mitochondrial gene *COI* datasets revealed deep genetic divergence among populations formerly considered as part of *H. pungens.*

These results agreed with the previous studies that revealed a deep genetic divergence, asymmetrical prezygotic and postzygotic reproductive isolation between the mealybug populations associated with different hosts from Argentina, and differential preference and performance on alternative hosts (Poveda-Martínez et al., 2019; Aguirre et al., 2016). The extension of our sampling effort to a wider geographic area throughout the native and non-native ranges, along with the use of a substantially large number of nuclear markers, allowed the confirmation of the results reported for the Argentine *Hypogeococcus* mealybugs (Poveda-Martínez et al., 2019) and new insights of the evolutionary history of *H. pungens sensu lato* . In effect, our present results supported the hypothesis that genetic differentiation throughout the entire range in South America has mainly been driven by divergent selection imposed by alternate host plants rather than by isolation through distance. In fact, Mantel correlation and multiple regression analyses of genetic distance matrices on geographic and host plant distance matrices revealed that the latter was a better predictor of genetic divergence among populations (Table 2).

In this context, our results pointed to host plant use as an important driver of cryptic divergence in *a priori* presumed polyphagous insects. Cryptic divergence has been recurrently observed in the form of host-associated species that, very likely, diverged as the result of adaptation to alternate host plants with little morphological divergence in several insect species (Matsubayashi, Kahono, & Katakura, 2011; Bagley, Sousa, Niemiller, & Linnen, 2017; Forbes et al., 2017; Zhang, Bass, Fernández, & Sharanowski, 2018; Driscoe et al., 2019).

Genetic divergence between the cactus feeding mealybugs native to Argentina and Paraguay, and those introduced to Australia from Argentina (ArPaAu-C) and cactus feeding mealybugs from southeastern Brazil (BrPr-C) exceeded that which would be expected for conspecific populations, as indicated by our species delimitation analyses. In this context, it is worth mentioning that besides the large geographic distance separating populations of these clades and the fact that both feed on Cactaceae, these two putative species fed on cactus that belong to different genera. Mealybugs of the ArAuPa-C clade fed on species of the genera *Cleistocactus* and *Harrisia* , whereas BrPr-C thrived on species of other genera like *Cereus* for native Brazilian mealybugs, and various genera for non-native Puerto Rican pest mealybugs (Table S1). It may be argued that divergent selection on alternate hosts and allopatry appeared as the more likely drivers of divergence. As a matter of fact, there is evidence that divergent selection caused by alternate hosts acting on ancestral fragmented populations is sufficient to produce genetic divergence and incidental speciation (Nosil et al., 2012; Duque-Gamboa et al., 2018; Doellman et al., 2019).

Our results also elucidated a certain degree of genetic heterogeneity within the ArAuPa-C clade, which included cactus feeders sampled in Argentina, Paraguay, and those introduced in Australia for biological control of cactus weeds. Both mtDNA and nuclear SNPs revealed a certain degree of genetic structuring within this clade. Several well differentiated mtDNA haplotypes were recorded at different locations of this

widespread clade, likely as the result of geographic isolation (Figure 2C). Also, the phylogenetic tree based on nuclear SNPs showed internal sub-clades concordant with the results depicted in the *COI* network. An evaluation of the external morphology of specimens collected on cacti in the same localities sampled for the genomic survey in Argentina revealed subtle morphological variation (Lucía Claps, University of Tucumán, Argentina, personal communication), suggesting that genetic heterogeneity in this clade may be indicative of incipient speciation rather than within species heterogeneity. These results also helped to end the debate of the controversial origin of the mealybugs used in the biological control program in Australia (and in South Africa) (Tomley & McFadyen, 1987; Williams, 1973; Hamon, 1984). Indeed, Australian *Hypogeococcus* mealybugs were genetically close to the cactus feeding samples collected in northwestern Argentina and Paraguay, but not, as suggested by other authors, to *Hypogeococcus festerianus* Lizer & Trelles, another cactophagous species inhabiting central-western Argentina, or *H. pungens* (Julien & Griffiths, 1999; Zimmermann, Pérez, Cuen, Mandujano, & Golubov, 2010; McFadyen, 2012).

Overall, our data provided evidence of two factors that influenced at least 48% (Table 2) of the genetic divergence between the mealybugs considered as part of the *H. pungens* species complex. Host plant associations seemed to be the primary force influencing genetic divergence, followed by the limited gene flow induced by isolation by distance. Still, these results should be interpreted with caution since other factors could also affect the distribution of genetic variation that remained unexplained. Local adaptation to different environmental conditions and/or ecological interactions with natural enemies or competitors may impose varying selective pressures in different geographic locations. A recent survey of parasitoids and hyperparasitoids associated with *Hypogeococcus* mealybugs in South America identified *Leptomastidea hypogeococci* Triapitsyn (Hymenoptera: Encyrtidae) as a widespread primary parasitoid, able to attack all members of the *H. pungens* complex, whereas *Anagyrus cachamai* Triapitsyn, Logarzo & Aguirre and *A. lapachosus* Triapitsyn, Aguirre & Logarzo (also encyrtid primary parasitoids from Argentina and Paraguay) were only found associated to *H. pungens sensu stricto* and the ArAuPr-C clade (Triapitsyn et al., 2018). The influence of such differential ecological interactions with natural enemies might be affecting patterns of genetic divergence in this species complex.

Results of phylogenetic and species delimitation analyses were not entirely congruent. Analysis with mtDNA visualized four major clades whereas genome wide SNPs allowed to detect five well supported clades (Figure 3). The main difference being that *Hypogeococcus* mealybugs from northeastern Brazil, Puerto Rico and United States (BrPRUS-AP clade) and *H. pungens sensu stricto* (Ar-A) were different species according to nuclear SNPs, while these clades appeared collapsed in the same group with mtDNA data (Figure 3). Such mito-nuclear inconsistencies have often been reported in insects (Weigand et al., 2017; Hinojosa et al., 2019). Even though the mtDNA has been useful to trace the evolutionary history in many species groups (Hebert, Penton, Burns, Janzen, & Hallwach, 2004; Ball & Armstrong, 2006), incomplete lineage sorting and past hybridization events may obscure species delimitation based only on mtDNA data, particularly in recently diverged taxa (Hickerson, Meyer, & Moritz, 2006; Hinojosa et al., 2019; Després, 2019). For instance, Moreyra et al., (2019) reported a mitogenomic study in a cluster of closely related cactophagous *Drosophila* spp. inhabiting the southern cone of South America and found that the evolutionary history inferred by mitogenomes is not completely concordant with the phylogeny depicted by nuclear genomes (Hurtado, Almeida, Revale, & Hasson, 2019). The authors proposed that either incomplete lineage sorting (ILS) and/or introgressive hybridization could account for the pattern observed. A word of caution is needed before arriving at definitive conclusions in the present study due to limitations of the mtDNA dataset, that consisted of only a few hundred base pairs of the mtDNA gene encoding *COI* . In contrast, the evolutionary history depicted by the nuclear genomic dataset may be considered more reliable since it consisted of more than one thousand widely distributed SNPs.

Phylogenetic trees based on either nuclear SNPs or mtDNA indicated that feeding on Amaranthaceae was the more likely ancestral condition. However, results yielded by mtDNA and SNPs datasets were not completely congruent concerning the evolution of host plant use; both cactus feeding species formed a derived monophyletic clade in the tree based on nuclear SNPs, while cactophagy appeared to have evolved twice in the tree obtained with mtDNA data. However, to unveil the ancestral host in these clades, and to evaluate

the evolution of host plant use and the biogeographic history of the genus in the continent, other well delimited cactophagous species of the genus *Hypogeococcus* from South America, such as *Hypogeococcus spinosus* Ferris and *H. festerianus* , should be included in an expanded analysis.

4.2 Puerto Rican cactus mealybugs derive from southeastern Brazilian cactus feeding populations

Our analyses, based on nuclear SNPs and mtDNA, allowed us to establish that the Puerto Rican cactus pest derives from a population similar to the southeastern Brazilian cactus feeding clade (Figure 2 and 3). In our previous study using only mtDNA and restricted sampling in the native range (Argentina), the Puerto Rican cactus pest clustered close to*H. pungens sensu stricto* (Poveda-Martínez et al., 2019). However, when we expanded the sampling to Paraguay and along the southeast and northeastern Atlantic coast of Brazil and extended the genetic sampling to genome-wide SNPs, we found that the Puerto Rican cactus pest fell into the same clade with populations from southeastern Brazil (Figure 2). Moreover, samples collected on Cactaceae in southern California shared the same and unique mtDNA haplotype with the Puerto Rican cactus pest. This finding is a warning of the threat that the presence of this pest represents to cactus diversity in the United States and Mexico, where cactus diversity is high. Since the first detection of the pest in Puerto Rico in 2005 (Segarra-Carmona et al., 2010), the mealybug now attacks half of the 14 native Puerto Rican cactus species (including three endemic and two endangered species) occurring in dry forests, causing large gall-like tissue deformations that often lead to high plant mortality (Carrera-Martínez et al., 2015; Triapitsyn et al., 2020). With the new record of the pest in California, the pest has spread beyond its current distribution range to potential cactus hosts throughout North America (including Mexico) and the Caribbean. Identification of the source of the cactus pest that invaded Puerto Rico and the southern United States is an important accomplishment since it may help to develop more specific biological control strategies aimed at protecting wild cactus from this mealybug. In classical biological control programs, the correct identification of the target species is a key issue for searching for natural enemies of the pest in its native area (Hoelmer & Kirk, 2005). Thus, untangling the evolutionary history of the *H. pungens* species complex did not only have taxonomic and systematic relevance, but from a practical perspective, it indicated that the design of biological control strategies (e.g., search for natural enemies) against the pest should focus on southeastern Brazilian cactus feeding mealybugs.

Our results showed that Amaranthaceae and/or Portulacaceae feeding mealybugs collected in Puerto Rico and continental United States were also derived from Amaranthaceae and/or Portulacaceae feeding*Hypogeococcus* , though, in this case, from northeastern Brazil (Figures 2 and 3). These findings suggested two invasion events for mealybugs of the *H. pungens* species complex into Puerto Rico and the continental United States, one involving cactus feeders and the other mealybugs feeding on Amaranthaceae and/or Portulacaceae. Comparisons of the levels of genetic diversity in the native and invasive ranges of these mealybugs supported the idea of the occurrence of founder events during the colonization of Puerto Rico and southern United States. In both cases, introduced populations showed lower levels of genetic diversity in both mtDNA and nuDNA than in the respective native ranges, suggesting that a reduced number of individuals were involved in each colonization process (Table 1). Many invasive species have been capable of thriving in novel environments despite the reduction of genetic variation as a consequence of founder events that may negatively impact fitness, survival, and evolutionary potential of the invasive populations (Tsutsui, Suarez, Holway, & Case, 2000; Logan, Minnaar, Keegan, & Clusella-Trullaset, 2020; Koch et al., 2020). In this vein, two putative species of the *H. pungens* complex have been successful invaders and continue to spread throughout the Caribbean and southern United States, threatening native species and cactus diversity despite the loss of genetic variation.

ACKNOWLEDGMENTS

We thank Andres Fernando Sanchez Restrepo and Nadia Jimenez for technical assistance, and Fabian Font for identification of the host plants. We appreciate the revision of Mayra Vidal who gave us important suggestions on the manuscript. Partial funding was obtained from USDA-APHIS Farm Bill 17-8130-0757-IA and 19-8130-0852-IA, and USDA-APHIS Biological Control Program 18-8130-0757-IA. Permits for fieldwork

in Brazil were provided by the SisGen (A9273F3) and the IBAMA (16BR022349/DF - 18BR027829/DF). Permits from Puerto Rico were provided by the U.S Fish & Wildlife Service (41522-16-003), and Department of Natural and Environmental Resources (OV-1617-15). D.P.M. is recipient of a PhD scholarship and M.B.A. of a postdoctoral fellowship both awarded by CONICET. E.H. is a member of CONICET Carrera del Investigador Cientifico.

REFERENCES

Aguirre, M. B., Diaz-Soltero, H., Claps, L. E., Saracho Bottero, A., Triapitsyn, S., Hasson, E., & Logarzo, G. A. (2016). Studies on the biology of *Hypogeococcus pungens* (sensu stricto)(Hemiptera: Pseudococcidae) in Argentina to aid the identification of the mealybug pest of Cactaceae in Puerto Rico. *Journal of Insect Science* ,*16* (1), 58.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* , 19(9), 1655-1664.

Anderson, C. J., Tay, W. T., McGaughran, A., Gordon, K., & Walsh, T. K. (2016). Population structure and gene flow in the global pest,*Helicoverpa armigera* . *Molecular Ecology, 25* (21), 5296-5311.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* , *17* (2), 81.

Bagley, R. K., Sousa, V. C., Niemiller, M. L., & Linnen, C. R. (2017). History, geography and host use shape genome wide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*).*Molecular Ecology* , *26* (4), 1022-1044.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* ,*3* (10).

Bakovic, V., Schuler, H., Schebeck, M., Feder, J. L., Stauffer, C., & Ragland, G. J. (2019). Host plant-related genomic differentiation in the European cherry fruit fly, *Rhagoletis cerasi*. *Molecular Ecology* ,*28* (20), 4648-4666.

Ball, S. L., & Armstrong, K. F. (2006). DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Canadian Journal of Forest Research* , 36(2), 337-350.

Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* , 16(1), 37-48.

Ben-Dov, Y. (1994). A systematic catalogue of the mealybugs of the world (Insecta: Homoptera: Coccoidea: Pseudococcidae and Putoidae) with data on geographical distribution, host plants, biology and economic importance. Intercept Limited.

Bouckaert, R., & Heled, J. (2014). DensiTree 2: Seeing trees through the forest. BioRxiv, 012401.

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., ... & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* , 10(4).

Brady, S. G., Gadau, J., & Ward, P. S. (2000). Systematics of the ant genus *Camponotus* (Hymenoptera: Formicidae): a preliminary analysis using data from the mitochondrial gene cytochrome oxidase I. In*Hymenoptera: evolution, biodiversity and biological control. Fourth International Hymenoptera Conference, held in Canberra, Australia, in January 1999.* (pp. 131-139). CSIRO Publishing.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis.*Molecular Biology and Evolution* , *29* (8), 1917-1932.

Bushnell, B. (2017). BBMap short read aligner, and other bioinformatic tools, version 37.33. (https://sourceforge.net/projects/bbmap/)

Carrera-Martinez, R., Aponte-Diaz, L., Ruiz-Arocho, J., & Jenkins, D. A. (2015). Symptomatology of infestation by *Hypogeococcus pungen* s: contrasts between host species. *Haseltonia ,2015* (21), 14-18.

Claps, L. E., & de Haro, M. E. (2001). Coccoidea (Insecta: Hemiptera) associated with Cactaceae in Argentina. *Journal of the Professional Association for Cactus Development* , 4: 77–83.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics* , 27(15), 2156-2158.

Despres, L. (2019). One, two or more species? Mitonuclear discordance and species delimitation. *Molecular Ecology* , *28* (17), 3845-3847.

Dinsdale, A., Cook, L., Riginos, C., Buckley, Y. M., & De Barro, P. (2010). Refined global analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae) mitochondrial cytochrome oxidase 1 to identify species level genetic boundaries. *Annals of the Entomological Society of America, 103* (2), 196-208.

Doellman, M. M., Schuler, H., Jean Saint, G., Hood, G. R., Egan, S. P., Powell, T. H., ... & Goughnour, R. B. (2019). Geographic and ecological dimensions of host plant-associated genetic differentiation and speciation in the *Rhagoletis cingulata* (Diptera: Tephritidae) sibling species group. *Insects* , 10(9), 275.

Driscoe, A. L., Nice, C. C., Busbee, R. W., Hood, G. R., Egan, S. P., & Ott, J. R. (2019). Host plant associations and geography interact to shape diversification in a specialist insect herbivore. *Molecular Ecology* , 28(18), 4197-4211.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* , *29* (8), 1969-1973.

Duque-Gamboa, D. N., Castillo-Cardenas, M. F., Hernandez, L. M., Guzman, Y. C., Manzano, M. R., & Toro-Perea, N. (2018). Mitochondrial DNA suggests cryptic speciation in *Prodiplosis longifila* Gagne (Diptera: Cecidomyiidae) associated with geographic distance and host specialization. *Bulletin of Entomological Research* , 108(6), 739-749.

Egan, S. P., Nosil, P., & Funk, D. J. (2008). Selection and genomic differentiation during ecological speciation: isolating the contributions of host association via a comparative genome scan of*Neochlamisus bebbianae* leaf beetles. *Evolution: International Journal of Organic Evolution* , *62* (5), 1162-1181.

Elfekih, S., Etter, P., Tay, W. T., Fumagalli, M., Gordon, K., Johnson, E., & De Barro, P. (2018). Genome-wide analyses of the *Bemisia tabaci* species complex reveal contrasting patterns of admixture and complex demographic histories. *PloS One* , *13* (1).

Eyer, P. A., Seltzer, R., Reiner-Brodetzki, T., & Hefetz, A. (2017). An integrative approach to untangling species delimitation in the*Cataglyphis bicolor* desert ant complex in Israel. *Molecular Phylogenetics and Evolution* , *115* , 128-139.

Ezard, T., Fujisawa, T., & Barraclough, T. G. (2009). Splits: species' limits by threshold statistics. *R package version* , *1* (11), r29. Available from: URL http://R-Forge.R-project.org/projects/splits/

Forbes, A. A., Devine, S. N., Hippee, A. C., Tvedte, E. S., Ward, A. K., Widmayer, H. A., & Wilson, C. J. (2017). Revisiting the particular role of host shifts in initiating insect speciation. *Evolution* ,*71* (5), 1126-1137.

Frichot, E., & Francois, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution,*6(8), 925-929.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & Francois, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* , 196(4), 973-983.

Funk, D. J., Nosil, P., & Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences, 103* (9),

3209-3213.

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software,* 22(7), 1-19.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* , 5(1), 184-186.

Granara de Willink, M. C. (1981). Nueva especie de *Hypogeococcus*Rau de Tucuman, Republica Argentina (Homoptera: Pseudococcidae).*Neotropica* , *27* , 61-65.

Hamon, A. B. (1984). A cactus mealybug, *Hypogeococcus festerianus*(Lizer y Trelles). *Florida (Homoptera: Coccoidea: Pseudococcidae). Entomology Circular, Division of Plant Industry, Florida Department of Agriculture and Consumer Services* , *263* .

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator* .*Proceedings of the National Academy of Sciences,* 101(41), 14812-14817.

Hickerson, M. J., Meyer, C. P., & Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space.*Systematic Biology* , *55* (5), 729-739.

Hinojosa, J. C., Koubinova, D., Szenteczki, M. A., Pitteloud, C., Dincă, V., Alvarez, N., & Vila, R. (2019). A mirage of cryptic species: genomics uncover striking mitonuclear discordance in the butterfly *Thymelicus sylvestris. Molecular Ecology, 28* (17), 3857-3868.

Hoelmer, K. A., & Kirk, A. A. (2005). Selecting arthropod biological control agents against arthropod pests: Can the science be improved to decrease the risk of releasing ineffective agents? *Biological Control, 34* (3), 255-264.

Hurtado, J. P., Almeida, F., Revale, S., & Hasson, E. (2019). Revised phylogenetic relationships within the *Drosophila buzzatii* species cluster (Diptera: Drosophilidae: Drosophila repleta group) using genomic data. *Arthropod Systematics and Phylogeny, 77* (2).

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., ... & Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research,27* (5), 768-777

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* , 11(1), 94.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* , 24(11), 1403-1405.

Julien M.H. Griffiths M.W. 1999. Biological control of weeds. A world catalogue of agent and their target weeds, 4th edn.CAB Publishing, Wallingford.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* , 90, 773– 795.

Koch, J. B., Dupuis, J. R., Jardeleza, M. K., Ouedraogo, N., Geib, S. M., Follett, P. A., & Price, D. K. (2020). Population genomic and phenotype diversity of invasive *Drosophila suzukii* in Hawaii.*Biological Invasions* , 22, 1753–1770.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets.*Molecular Biology and Evolution* , *33* (7), 1870-1874.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic Biology* , 63(4), 534-542.

Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* , 10(5), 831–844

Leigh, J. W., & Bryant, D. (2015). Popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* , 6(9), 1110-1116.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25* (16), 2078-2079.

Logan, M. L., Minnaar, I. A., Keegan, K. M., & Clusella-Trullas, S. (2020). The evolutionary potential of an insect invader under climate change. *Evolution* , 74(1), 132-144.

Loxdale, H. D., & Harvey, J. A. (2016). The 'generalism'debate: misinterpreting the term in the empirical literature focusing on dietary breadth in insects. *Biological Journal of the Linnean Society ,119* (2), 265-282.

Malka, O., Santos-Garcia, D., Feldmesser, E., Sharon, E., Krause-Sakate, R., Delatte, H., ... & Seal, S. (2018). Species-complex diversification and host-plant associations in *Bemisia tabaci:* A plant-defence, detoxification perspective revealed by RNA-Seq analyses. *Molecular Ecology* , *27* (21), 4241-4256.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* , *27* (2 Part 1), 209-220.

Matsubayashi, K. W., Kahono, S., & Katakura, H. (2011). Divergent host plant specialization as the critical driving force in speciation between populations of a phytophagous ladybird beetle. *Journal of Evolutionary Biology, 24* (7), 1421-1432.

McFadyen, R. E., & Tomley, A. J. (1978). Preliminary indications of success in the biological control of Harrisia cactus (*Eriocereus martinii* Lab.) in Queensland. In *Proceedings of the First Conference of the Council of Australian Weed Science Societies* held at National Science Centre, Parkville, Victoria, Australia, 12-14 April 1978 (pp. 108-112). Council of Australian Weed Science Societies.

McFadyen, R. E., & Tomley, A. J. (1981). Biological control of harrisia cactus, *Eriocereus martinii,* in Queensland by the mealy bug,*Hypogeococcus festerianus.* In *Proceedings of the 5th International Symposium on Biological Control of Weeds* . (pp. 589-594). Commonwealth Scientific and Industrial Research Organization..

McFadyen R. E. (2012). *Harrisia (Eriocereus) martinii* (Labour.) Britton—Harrisia cactus *Acanthocereus tetragonu* s (L.) Hummelink—sword pear, pp. 274–281. InJulien M McFadyen R Cullen J (eds), *Biological control of weeds in Australia* . CSIRO Publishing, Collingwood, Australia.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *2010 gateway computing environments workshop (GCE)* (pp. 1-8).

Moreyra, N. N., Mensch, J., Hurtado, J., Almeida, F., Laprida, C., & Hasson, E. (2019). What does mitogenomics tell us about the evolutionary history of the *Drosophila buzzatii* cluster (repleta group)?.*PloS One* , 14(11).

Newbold, T., Hudson, L. N., Hill, S. L., Contu, S., Lysenko, I., Senior, R. A., ... & Day, J. (2015). Global effects of land use on local terrestrial biodiversity. *Nature* , 520(7545), 45-50.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* , 32(1), 268-274.

Nosil, P., Gompert, Z., Farkas, T. E., Comeault, A. A., Feder, J. L., Buerkle, C. A., & Parchman, T. L. (2012). Genomic consequences of multiple speciation processes in a stick insect. *Proceedings of the Royal Society B: Biological Sciences* , *279* (1749), 5058-5065.

Nosil, P. (2012). Ecological speciation. Oxford University Press.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. A. S. S. (2007). The vegan package.*Community Ecology Package* , *10* , 631-637.

Orsucci, M., Audiot, P., Nidelet, S., Dorkeld, F., Pommier, A., Vabre, M., ... & Streiff, R. (2018). Transcriptomic response of female adult moths to host and non-host plants in two closely related species. BMC evolutionary biology, 18(1), 145.

Paterson, I. D., Hoffmann, J. H., Klein, H., Mathenge, C. W., Neser, S., & Zimmermann, H. G. (2011). Biological control of cactaceae in South Africa. *African Entomology* , 19(2), 230-246.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., ... & Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology, 55* (4), 595-609.

Poveda-Martinez, D., Aguirre, M. B., Logarzo, G., Calderon, L., de la Colina, A., Hight, S., ... & Hasson, E. (2019). Untangling the*Hypogeococcus pungens* species complex (Hemiptera: Pseudococcidae) for Argentina, Australia, and Puerto Rico based on host plant associations and genetic evidence. *PloS One* , 14(7).

Powell, T. H., Forbes, A. A., Hood, G. R., & Feder, J. L. (2014). Ecological adaptation and reproductive isolation in sympatry: genetic and phenotypic evidence for native host races of *Rhagoletis pomonella. Molecular Ecology, 23* (3), 688-704.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* , 155(2), 945-959.

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ,* 2, e431.

Rambaut, A., Suchard, M.A., Xie, D., & Drummond, A.J. (2014). Tracer v1.6. [Software]. Available: http://beast.bio.ed.ac.uk/Tracer BEAST

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Hohna, S., ... & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology, 61* (3), 539-542.

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S., & Sanchez-Gracia, A. (2017). DnaSP v6: DNA sequence polymorphism analysis of a large datasets. *Molecular Biology and Evolution,* 34, 3299-3302

Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ* , *3* , e1106

Ryan, S. F., Lombaert, E., Espeset, A., Vila, R., Talavera, G., Dincă, V., ... & Li, Y. (2019). Global invasion history of the agricultural pest butterfly *Pieris rapae* revealed with genomics and citizen science. *Proceedings of the National Academy of Sciences, 116* (40), 20015-20024.

Schoonhoven, L. M., Van Loon, B., van Loon, J. J., & Dicke, M. (2005).*Insect-plant biology* . Oxford University Press on Demand.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., & Flook, P. (1994). Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America* ,*87* (6), 651-701.

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence.*Systematic Zoology* , *35* (4), 627-632.

Stouthamer, R., Rugman-Jones, P., Thu, P. Q., Eskalen, A., Thibault, T., Hulcr, J., ... & Lin, C. (2017). Tracing the origin of a cryptic invader: phylogeography of the *Euwallacea fornicatus* (Coleoptera: Curculionidae: Scolytinae) species complex. *Agricultural and Forest Entomology* , *19* (4), 366-375.

Tomley, A. J., & McFadyen, R. E. (1984). Biological control of Harrisia cactus, *Eriocereus martinii,* in central Queensland by the mealybug, *Hypogeococcus festerianus* , nine years after release. In*Proceedings of VI International Symposium of Biological Control of Weeds. Agriculture Canada, Vancouver* (pp. 843-847).

Triapitsyn, S. V., Aguirre, M. B., Logarzo, G. A., Hight, S. D., Ciomperlik, M. A., Rugman-Jones, P. F., & Verle Rodrigues, J. C. (2018). Complex of primary and secondary parasitoids (Hymenoptera: Encyrtidae and Signiphoridae) of *Hypogeococcus* spp. mealybugs (Hemiptera: Pseudococcidae) in the New World. Florida Entomologist,*101* (3), 411-434.

Triapitsyn, S. V., Hight, S. D., Logarzo, G. A., Aguirre, M. B., Verle Rodrigues, J. C., Trjapitzin, V. A., ... & Rodríguez Reyes, Y. (2020). Natural enemies of the Harrisia cactus mealybug and other*Hypogeococcus* species (Hemiptera: Pseudococcidae) in Puerto Rico: identification and taxonomic notes on primary and secondary parasitoids.*Neotropical Entomology,* Online First: https://doi.org/10.1007/s13744-019-00754-w, 1-23.

Tsutsui, N. D., Suarez, A. V., Holway, D. A., & Case, T. J. (2000). Reduced genetic variation and the success of an invasive species.*Proceedings of the National Academy of Sciences* , *97* (11), 5948-5953.

Vidal, M. C., Quinn, T. W., Stireman III, J. O., Tinghitella, R. M., & Murphy, S. M. (2019). Geography is more important than host plant use for the population genetic structure of a generalist insect herbivore.*Molecular Ecology* , *28* (18), 4317-4334.

Weigand, H., Weiss, M., Cai, H., Li, Y., Yu, L., Zhang, C., & Leese, F. (2017). Deciphering the origin of mito-nuclear discordance in two sibling caddisfly species. *Molecular Ecology* , 26(20), 5705-5715.

Williams, D. J. (1973). Two cactus-feeding mealybugs from Argentina (Homoptera, Coccoidea, Pseudococcidae). *Bulletin of Entomological Research* , *62* (4), 565-570.

Winter, S., Friedman, A. L., Astrin, J. J., Gottsberger, B., & Letsch, H. (2017). Timing and host plant associations in the evolution of the weevil tribe *Apionini* (Apioninae, Brentidae, Curculionoidea, Coleoptera) indicate an ancient co-diversification pattern of beetles and flowering plants. *Molecular Phylogenetics and Evolution* ,*107* , 179-190.

Zhang, J. J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics,* 29, 2869–2876.

Zhang, Y. M., Bass, A. I., Fernandez, D. C., & Sharanowski, B. J. (2018). Habitat or temporal isolation: Unraveling herbivore–parasitoid speciation patterns using double digest RADseq. *Ecology and Evolution* , *8* (19), 9803-9816.

Zimmermann, H. G., Perez Sandi Cuen, M., Mandujano, M. C., & Golubov, J. (2010). The South American mealybug that threatens North American cacti. *Cactus and Succulent Journal* , *82* (3), 105-107.

DATA ACCESSIBILITY STATEMENT

Raw sequence data are available at the National Centre for Biotechnology Information (NCBI), under the BioProject accession number PRJNA593002. Reference genome is available as FuEDEI_HPun_1.1.fa in NCBI under accession number: JAAOIU000000000. New mtDNA haplotypes were deposited in GenBank under access numbers: MT138921 - MT138931.

AUTHOR CONTRIBUTION

D.P.M., M.B.A., G.L., S.D.H., E.H. designed the project. D.P.M., M.B.A., performed research and analysed data, M.B.A., E.H., G.L., S.D.H., M.V.D., S.V.T., H.D.S. made field collection. G.L., S.D.H., H.D.S.,

obtained funding. D.P.M., M.B.A., E.H. wrote the first draft and all co-authors contributed equally to the revisions.

FIGURE CAPTIONS

Figure 1. Locations of mealybug collection sites in both the native and non-native range of *Hypogeococcus pungens* species complex. Green circles represent Cactaceae host plants where mealybugs were collected; purple circles represent Amaranthaceae host plants, while yellow circles represent Portulacaceae host plants. A and B correspond with the native range distribution, while C, D, E and F correspond with the non-native range distribution of the species complex.

Figure 2. Clustering analyses using sNMF (A) and DAPC (B) methods based on 1,707 SNPs from 141 *Hypogeococcus pungens*species complex specimens and neighbor joining network based on mtDNA data (C). Visual representation of five clusters (K= 5) along with membership probability of each individual to the corresponding cluster. Color codes in both clustering analyses and in the neighbor joining network are the same: yellow: Ar-A: specimens from Argentina feeding on Amaranthaceae; green: BrPRUS-AP: specimens from northern Brazil, Puerto Rico and the United States feeding on Amaranthaceae and/or Portulacaceae; red: ArPaAu-C: specimens from Argentina, Paraguay and Australia feeding on Cactaceae; light blue: BrPR-C: specimens from southern Brazil and Puerto Rico feeding on Cactaceae and blue: Br-A: specimens from southern Brazil feeding on Amaranthaceae.

Figure 3. Phylogenetic hypotheses based on SNPs and mtDNA for *Hypogeococcus pungens* species complex. Phylogenetic trees were inferred by ML and BI based on 1,707 SNPs (A) and mtDNA haplotypes (B). Species tree reconstructed using coalescent method in SNAPP based on the dataset of 1,679 SNPs (C). Mealybugs feeding on Cactaceae host plant: clade ArPaAu-C, mealybugs from Argentina, Paraguay and Australia, and mealybugs from southeastern Brazil and Puerto Rico (clade BrPR-C). Mealybugs feeding on Amaranthaceae and Portulacaceae host plant: Br-A, mealybugs from southeastern Brazil; clade Ar-A, *H. pungens sensu stricto* ; clade clade BrPRUS-AP, mealybugs from northeastern Brazil, Puerto Rico and the United States.

SUPPORTING INFORMATION

Table S1. Host plant use and geographic origin of*Hypogeococcus pungens* species complex samples.

Table S2. Number of loci identified after each quality filtering step.

Table S3. Pairwise Fst estimates for *Hypogeococcus pungens* species complex populations sampled in native and invaded areas.

Table S4. Pairwise Fst estimates for the groups defined by cluster analysis.

Table S5. Species delimitation scenarios for*Hypogeococcus pungens* species complex.

Figure S1 . Outcomes of single locus species delimitation analyses.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ORCID

Daniel Poveda-Martinez*https://orcid.org/0000-0003-3657-8421*

Maria Belen Aguirre*https://orcid.org/0000-0002-3773-8805*

Stephen D. Hight*https://orcid.org/0000-0003-0832-394X*

Serguei Triapitsyn*http://orcid.org/0000-0002-5086-7847*

Marcelo Diniz Vitorino*https://orcid.org/0000-0002-9537-6275*

Esteban Hasson*https://orcid.org/0000-0003-3326-9877*

Hosted file

Table_1.docx available at https://authorea.com/users/318612/articles/448543-species-complex-diversification-by-host-plant-use-in-an-herbivorous-insect-the-source-of-puerto-rican-cactus-mealybug-pest-and-implications-for-biological-control

Hosted file

Table_2.docx available at https://authorea.com/users/318612/articles/448543-species-complex-diversification-by-host-plant-use-in-an-herbivorous-insect-the-source-of-puerto-rican-cactus-mealybug-pest-and-implications-for-biological-control