Received: 16 February 2017 | Accepted: 31 August 2017

DOI: 10.1111/eva.12559

ORIGINAL ARTICLE

WILEY Evolutionary Applications Open Access

Assessing the potential of genotyping-by-sequencing-derived single nucleotide polymorphisms to identify the geographic origins of intercepted gypsy moth (*Lymantria dispar*) specimens: A proof-of-concept study

Sandrine Picq[1,2] | Melody Keena[3] | Nathan Havill[3] | Don Stewart[1] | Esther Pouliot[1] | Brian Boyle[2] | Roger C. Levesque[2] | Richard C. Hamelin[2,4] | Michel Cusson[1,2]

[1]Laurentian Forestry Centre, Natural Resources Canada, Quebec City, QC, Canada

[2]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec City, QC, Canada

[3]USDA Forest Service, Northern Research Station, Northeastern Center for Forest Health Research, Hamden, CT, USA

[4]Department of Forest Sciences, Faculty of Forestry, The University of British Columbia, Vancouver, BC, Canada

Correspondence
Sandrine Picq and Michel Cusson, Laurentian Forestry Centre, Natural Resources Canada, Quebec City, QC, Canada.
Emails: sandrine.picq@gmail.com and michel.cusson@canada.ca

Funding information
Genomics Research and Development Initiative; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 171350-2012; Genomic Applications Partnership Program, Grant/Award Number: #244DIA; Large Scale Applied Research Program, Grant/Award Number: #182IAS

Abstract

Forest invasive alien species are a major threat to ecosystem stability and can have enormous economic and social impacts. For this reason, preventing the introduction of Asian gypsy moths (AGM; *Lymantria dispar asiatica* and *L. d. japonica*) into North America has been identified as a top priority by North American authorities. The AGM is an important defoliator of a wide variety of hardwood and coniferous trees, displaying a much broader host range and an enhanced dispersal ability relative to the already established European gypsy moth (*L. d. dispar*). Although molecular assays have been developed to help distinguish gypsy moth subspecies, these tools are not adequate for tracing the geographic origins of AGM samples intercepted on foreign vessels. Yet, this type of information would be very useful in characterizing introduction pathways and would help North American regulatory authorities in preventing introductions. The present proof-of-concept study assessed the potential of single nucleotide polymorphism (SNP) markers, obtained through genotyping by sequencing (GBS), to identify the geographic origins of gypsy moth samples. The approach was applied to eight laboratory-reared gypsy moth populations, whose original stocks came from locations distributed over the entire range of *L. dispar*, comprising representatives of the three recognized subspecies. The various analyses we performed showed strong differentiation among populations ($F_{ST} \geq 0.237$), enabling clear distinction of subspecies and geographic variants, while revealing introgression near the geographic boundaries between subspecies. This strong population structure resulted in 100% assignment success of moths to their original population when 2,327 SNPs were used. Although the SNP panels we developed are not immediately applicable to contemporary, natural populations because of distorted allele frequencies in the laboratory-reared populations we used, our results attest to the potential of genomewide SNP markers as a tool to identify the geographic origins of intercepted gypsy moth samples.

KEYWORDS
genotyping-by-sequencing, invasive species, laboratory populations, *Lymantria dispar*, population assignment, SNP filtering

1 | INTRODUCTION

Invasive alien insects represent a major threat for biodiversity and ecosystem stability, and can have considerable economic and social impacts (Bradshaw et al., 2016; Kenis et al., 2009). The European gypsy moth, *Lymantria dispar dispar* Linnaeus, is a perfect example of an invasive alien species that is responsible for severe tree growth losses in its new habitat (Bradshaw et al., 2016). Since its accidental introduction from Europe into eastern North America in the late 1860s (Pogue & Schaefer, 2007), the European gypsy moth (EGM) has caused billions of dollars in losses for the forest industry and urban communities, and has required important investments in pest management (Bradshaw et al., 2016). In addition, this moth has altered biodiversity in its new habitat by contributing to a decline in oak populations in eastern North America (Morin & Liebhold, 2016). This severe impact of EGM can be explained by the wide variety of hardwood and coniferous trees defoliated by its larvae (Liebhold et al., 1995) and by periodic population irruptions, leading to outbreaks that cover large areas. Fortunately, range expansion of the gypsy moth in North America has been limited by the inability of EGM females to fly (Pogue & Schaefer, 2007). Dispersal is accomplished through crawling of caterpillars from tree to tree or larval ballooning, that is, aerial dispersal using silk, and is therefore limited to short distances (<100 m) (Lance & Barbosa, 1982; Nickason, 2001; Weseloh, 1997). Nevertheless, long-distance displacements have occurred repeatedly as a result of accidental transportation of egg masses on firewood, household goods and vehicles (Bigsby, Tobin, & Sills, 2011). Consequently, EGM has spread at a rate of 3 to 29 km/year since its introduction (Tobin, Liebhold, & Anderson Roberts, 2007) and is now considered established on a territory ranging from Quebec to North Carolina on the east coast, and inland to Wisconsin (Tobin, Bai, Eggen, & Leonard, 2012). EGM is currently considered one of the most destructive non-native insects in eastern North America (Aukema et al., 2011).

Two Asian subspecies of *L. dispar* have been described and both are referred to as “Asian gypsy moth”: *L. dispar asiatica* Vnukovskij, present in continental Asia, and *L. dispar japonica* Motschulsky, distributed across the Japanese archipelago (Pogue & Schaefer, 2007; Wu et al., 2015). The Asian gypsy moth (AGM) is considered a greater threat than its European counterpart to North America’s forests due to the strong flight capability of its females and its broader host range (Pogue & Schaefer, 2007). AGM is not established in North America, but egg masses and adult moths are recurrently intercepted during foreign vessel inspections in North American ports, and occasional escapees have been removed through major eradication campaigns (APHIS-USDA, 2006, 2016). Given that both AGM subspecies can interbreed with *L. d. dispar* (M. Keena, *unpublished data*), the escape and establishment of AGM in North America could lead to the introgression of traits such as strong female flight capacity and extended host range into North American EGM populations. For example, 16%–33% of the female progeny obtained from crosses between AGM and their North American counterpart are capable of strong flight (Keena, Grinberg, & Wallner, 2007). Thus, the establishment of AGM could result in accelerated spread and more severe outbreaks. In this context, the accurate identification of moths and egg masses intercepted on foreign vessels is critical if we are to avoid AGM establishment and introgression of undesirable traits into North American gypsy moth populations.

Distinguishing moths of each *L. dispar* subspecies using morphological characters has proven to be a nontrivial task, and gypsy moth egg masses cannot be visually identified at the subspecies level. For these reasons, several molecular diagnostic tools aimed at separating AGM from EGM (both North American and European populations of *L. dispar dispar*) have been developed (e.g., Bogdanowicz, Schaefer, & Harrison, 2000; deWaard et al., 2010; Garner & Slavicek, 1996; Pfeifer, Humble, Ring, & Grigliatti, 1995), including a qPCR-based suite of assays designed by our group (Stewart et al., 2016). However, molecular tools aimed at identifying the geographic origins of intercepted samples are still lacking; tracing the origins of such samples is critical to negotiations undertaken by Canadian and American regulatory authorities with trading partners to help prevent future introductions.

The advent of next-generation sequencing (NGS) has greatly facilitated the development of diagnostic single nucleotide polymorphism (SNP) markers, including for nonmodel organisms. Such NGS-derived SNPs have also been shown to provide enhanced resolution relative to classical markers (e.g., microsatellites), as in the identification of the geographic origins of individuals (Larson et al., 2014; Puckett & Eggert, 2016). Here we present the results of a proof-of-concept study aimed at assessing the potential of genotyping-by-sequencing (GBS)-derived SNPs to trace the geographic source of gypsy moth samples. In view of the proof-of-concept nature of our work and difficulties in rapidly obtaining fresh gypsy moth samples from the field during periods of low population densities, our study focused on eight laboratory-reared gypsy moth populations whose original stocks came from locations distributed over the entire range of *L. dispar*, comprising representatives of the three recognized subspecies.

In the present work, the SNPs obtained by GBS (Mascher, Wu, Amand, Stein, & Poland, 2013) were first submitted to both classical and more recent analytical approaches used in the field of population genomics (linkage disequilibrium network analysis, population structure analysis and F_{ST} calculations) to assess their usefulness in discriminating our gypsy moth populations. Then, the success of these SNPs in assigning moths to their original population was evaluated using two different methods: a multivariate approach, discriminant analysis of principal components (DAPC; Jombart, Devillard, & Balloux, 2010) and a Bayesian approach implemented in the Genetic Stock Identification software gsi_sim (Anderson, Waples, & Kalinowski, 2008). Finally, we evaluated the impact of modifying a SNP filtering parameter (e.g., linkage disequilibrium filtering) on the outcome of our analyses. Although the SNP panels we developed using the aforementioned assignment methods are not immediately applicable to contemporary, wild gypsy moth populations, our results attest to the potential of genomewide SNP markers as a tool to identify the geographic origins of intercepted gypsy moth samples.

2 | MATERIALS AND METHODS

2.1 | Insect material and DNA extraction

In view of the challenge associated with collecting fresh gypsy moth samples at various locations over this species' vast geographic range, we opted for the use of specimens derived from material collected in the context of earlier studies (Keena et al., 2007; Keena, Côté, Grinberg, & Wallner, 2008; H. Nadel, *unpublished data*) and subsequently maintained in the laboratory as distinct, population-specific colonies (see Figure 4 for a map showing collection locations). Such samples had the advantage of being readily available and well characterized with respect to female flight capability, two features that weighed heavily in our decision to use this material and in our assessment of its suitability for a proof-of-concept study. This material also provided an opportunity to assess the impact of laboratory rearing, over several generations, on genetic diversity.

The laboratory populations we used were established using 4 to 58 egg masses per population collected in the field 6 to 25 years ago (see Table 1). At each generation in the laboratory, 100 egg masses/population were chosen at random from all those produced by the previous generation (note: in the laboratory, the gypsy moth has a generation about every 8 months and each mated female lays a single egg mass containing 500–2,000 eggs). From these egg masses, 400 to 500 eggs were randomly selected and reared to the adult stage. Then, ~150 adult pairs/population were formed for mating to produce the next generation. Given that the date of adult emergence from pupae can vary, adults selected to produce eggs for the next generation were sampled to provide an even coverage of emergence dates. Although the populations used in our study had been bred in the laboratory for several generations, they appeared to have maintained their original biological and behavioural attributes. Of course, these insects have unavoidably undergone some adaptation to laboratory rearing, but every effort was made to limit, as much as possible, the loss of genetic diversity relative to the original field samples.

For the work presented here, we took a random sample of 12 specimens (six females and six males; Table 1) from each laboratory population, for a total of 96 moths. The number of individuals we selected from each population may be viewed as an approximation of the average number of gypsy moths caught in pheromone traps during endemic periods (Streifel, 2016); this value was deemed sufficient for the purpose of our study, given that accurate estimates of population differentiation and genetic diversity can be obtained with small samples (4–8 individuals) when many markers are considered (i.e., >1,000 SNPs; Willing, Dreyer, & van Oosterhout, 2012; Nazareno, Bemmels, Dick, & Lohmann, 2017). Specimens were assigned to one of the three recognized *L. dispar* subspecies (*L. d. dispar*, *L. d. asiatica* and *L. d. japonica*), using the criteria of Pogue and Schaefer (2007).

Before DNA extraction, heads and thoraces of adult moths were frozen in liquid nitrogen and ground using a Retsch MM 200 mixer mill (Retsch technology, Haan, Germany). DNA was extracted using the DNeasy 96 Blood & Tissue Kit (Qiagen, Carlsbad, CA, USA), according to the manufacturer's instructions, with the exception of an additional RNase A treatment before the addition of buffer AL/ethanol. DNA concentration and purity of the extracts were assessed using a NanoDrop 8000 spectrophotometer (Thermo Scientific, Waltham, MA, USA).

2.2 | Genotyping-by-sequencing library construction and sequencing

Prior to library construction, samples were diluted to 20 ng/μl. To prepare a reduced-representation library for sequencing, we used a modified genotyping-by-sequencing (GBS) protocol where two restriction enzymes (PstI/MspI) and a *Y*-adapter are employed (Mascher et al., 2013). To ensure sufficient read depth, the library was sequenced on three P1v3 chips using HiQ reagents on an Ion Proton sequencer (Thermo Scientific, Waltham, MA, USA). GBS library construction and sequencing were carried out at the Plate-forme d'analyses génomiques of the Institut de Biologie Intégrative et des Systèmes (IBIS) at Université Laval (Quebec City, QC, Canada).

2.3 | Bioinformatic analysis and genotyping

Prior to analysis, read quality was assessed using the FastQC software (Andrews, 2016). SNP calling was performed without a reference genome, using the Universal Network Enabled Analysis Kit (UNEAK) pipeline (Lu et al., 2013) with default parameter settings: minimum tag count $c = 5$, error tolerance rate in the network filter $e = 0.03$ and minimum minor allele frequency mnMAF = 0.05. Briefly, UNEAK retains all reads containing a barcode and a restriction enzyme cut site, in addition to being devoid of missing data in the first 64 bp after the barcode. Reads are then clustered into tags (i.e., reads displaying 100% identity), and only tags with $c \geq 5$ are retained. Then, networks of tags differing by one bp are built. In these different networks, tags with read counts corresponding to 3% ($e = 0.03$) or less of read counts from the adjacent tags are considered errors. The edges connecting the "error" tags to "real" tags are then sheared, thus dividing networks into subnetworks or decreasing the number of tags present in networks. At the end of the process, only tag-pair networks are retained as potential SNPs; networks with multiple tags are discarded.

2.4 | SNP filtering

Among SNPs identified by UNEAK, we retained those genotyped in ≥80% of individuals and present in at least seven populations of eight (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013). Then, loci presenting

TABLE 1 Description of populations used in the present study, along with intrapopulation diversity indices

				Origin								
Population[a]	*N*[b]	No. egg mass	Collection date	Country	Region	Latitude	Longitude	Subspecies[c]	Ho[d]	He[e]	Fis[f]	Ne[g]
JA†	12 (12)	>3	2005	Japan[h]	Iwate	39.66°N	141.43°E	*japonica*	0.127	0.150	0.148*	19.2 (18.3–20.2)
CB§	12 (12)	30	2011	China	Beijing	39.53°N	116.23°E	*asiatica*	0.153	0.169	0.095*	28.2 (26.6–29.9)
MG†	12 (11)	Unknown	Early 1990s	Mongolia[h]	–	46.41°N	103.13°E	*asiatica*	0.143	0.155	0.077*	31.5 (29.3–34.0)
RM§	12 (12)	20	1992	Russia	Primorski	44.10°N	113.15°E	*asiatica*	0.161	0.202	0.202*	98.4 (85.7–115.2)
RB§	12 (12)	30	1992	Russia	Krasnoyarsk	54.30°N	91.18° E	*asiatica*	0.170	0.197	0.138*	139.5 (116.0–174.0)
LJ§	12 (10)	47	1994	Lithuania	Kuzsin Nezijos	55.31°N	21.06°E	*dispar*	0.128	0.184	0.306*	106.3 (86.4–137.7)
KG§	12 (12)	58	1997	Greece	Macedonia	41.00°N	24.25°E	*dispar*	0.128	0.150	0.145*	27.3 (25.6–29.1)
UC§	12 (10)	12	1994	United States	Connecticut	41.25°N	73.00°W	*dispar*	0.106	0.135	0.215*	2.8 (2.7–2.8)

[a] Populations maintained as pure laboratory rearings by M. Keena (§) and H. Nadel (†).

[b] No. of moths from which DNA was extracted (No. of samples left for analysis after data filtering).

[c] Using the identification criteria of Pogue and Schaefer (2007).

[d] Mean observed heterozygosity.

[e] Mean expected heterozygosity.

[f] Inbreeding coefficient.

[g] Estimate of contemporary effective population size; in parentheses, 95% confidence interval.

[h] Exact locality unknown; the geographic coordinates correspond to the centre of the region/country.

**p*-value < .001.

a mean coverage per individual ≤10 reads were also discarded. SNPs with an average minor allele frequency (MAF) < 0.05 across populations were also discarded (very low-frequency SNPs create biases in quantifying genetic differentiation F_{ST} and identifying outlier SNPs; Roesti, Salzburger, & Berner, 2012).

Paralogous sequences differing by one nucleotide can be erroneously identified as alleles of a same locus by the UNEAK pipeline. Such misidentification will result in uninformative false SNPs for which almost all individuals appear heterozygous. This problem of confusing paralogs with allelic variants is limited inasmuch as two or more mutations are sufficient for UNEAK to distinguish paralogous loci. Thus, SNPs originating from the overmerging of paralogous sequences were excluded by removing markers showing observed heterozygosity >0.50 (Hohenlohe, Amish, Catchen, Allendorf, & Luikart, 2011). Extreme deviation from Hardy–Weinberg equilibrium ($p < .01$) in three or more populations was used to remove SNPs with important genotyping errors (Teo, Fry, Clark, Tai, & Seielstad, 2007).

Highly linked SNPs could introduce bias in analyses requiring independence of loci, for example, outlier SNP detection analysis and model-based methods to describe population structure. For this reason, we identified pairs of SNPs displaying high linkage disequilibrium (LD; $r^2 > 0.80$) in at least three populations and eliminated the SNP, within the pair, showing the most missing data. This LD filtering criterion led to 35 pairs of SNPs being identified as highly linked. The presence of such SNPs gave us an opportunity to assess their influence on the results of our analyses. Thus, all analyses presented here, with the exception of assignment tests, were run on data sets with and without high-LD SNPs.

Impacts of the different filtering steps on SNP counts are reported in Table 2. Missing data filtering and LD calculations were carried out using VCFtools (Danecek et al., 2011), whereas the other filtering procedures were conducted using an in-house R script. The resulting VCF file was converted to file formats suitable for each subsequent analysis using PGDSpider v2.0.9.2 (Lischer & Excoffier, 2012).

TABLE 2 Number of SNPs retained after each sequential filtering step

SNP identification step	Filtered SNPs	Remaining SNPs
UNEAK output		58,309
Coverage		
> 80% of specimens and 7 strains of 8	54,997	3,312
No. reads per locus > 10	510	2,802
Frequency		
Minor allele frequency > 0.05	231	2,571
Hobs < 0.5	148	2,423
Hardy–Weinberg equilibrium	61	2,362
Linkage disequilibrium	35	2,327
Outliers	133	2,194

2.5 | Linkage disequilibrium network analysis

The analysis of linkage disequilibrium (LD), that is, the nonrandom association of alleles from different loci, can reveal various evolutionary processes in population genomic data sets, including local adaptation and geographic structure (Kemppainen et al., 2015). To explore LD patterns in our data set, we used linkage disequilibrium network analysis (LDna) as implemented in the *LDna* R package (Kemppainen et al., 2015). This analytical procedure, which identifies groups of loci in high LD, does not require knowledge of locus positions within the genome and is therefore applicable to species without a reference genome. In addition, LDna explores LD across the entire genome, thus generating information about LD among widely scattered loci, which classical LD analysis does not do.

Using a matrix of pairwise LD estimates among loci, LDna produces a tree where branches represent loci and/or clusters of loci and the joining of branches represents individual loci and/or clusters of loci that merge at a particular LD threshold. LDna analysis is based on the hypothesis that clusters remaining separate across a wide range of LD thresholds represent different genetic signals in the data. A change in LD when two branches merge is quantified by λ, for which high values indicate the merger of large clusters and/or clusters with high pairwise LD values. Any cluster displaying a λ value exceeding the median of all λ values by a user-defined multiple φ of the median absolute deviation and containing at least |E|min edges (also user-defined) is considered an outlier cluster. Outlier clusters that do not have any other outlier clusters nested within them are defined as single-outlier clusters (SOCs). For our data set, we used the *dudi.pca* function of *adegenet* package (Jombart et al., 2010) to carry out principal component analysis (PCA) on loci of each identified SOCs to determine whether geographic structure or other evolutionary phenomena were responsible for the observed LD pattern.

Given that LDna is sensitive to missing data, rare alleles and loci displaying heterozygosity > 0.5 (Picq, McMillan, & Puebla, 2016), the analyses were conducted on the filtered data set. Prior to carrying out LDna analysis, LD was measured as the squared pairwise correlation coefficient (r^2) between loci using VCFtools (Danecek et al., 2011) and the matrix of pairwise LD values was generated using an in-house R script.

2.6 | Detecting outlier SNPs

Single nucleotide polymorphisms with extreme F_{ST} values can greatly affect population differentiation estimates and phylogenetic inferences (Luikart, England, Tallmon, Jordan, & Taberlet, 2003). To accurately identify outlier SNPs, we used a combination of methods, as suggested by Pérez-Figueroa, García-Pereira, Saura, Rolán-Alvarez, and Caballero (2010). First, BayeScan 2.0 (Foll & Gaggiotti, 2008) was used for a Bayesian analysis. This software accommodates differences in population effective size and immigration rate among subpopulations, and can take into account uncertainty about allele frequency resulting from small sample sizes. The program was run with default parameters, and SNPs with q-value ≤0.05 were considered outliers

(q-value is the false discovery rate analogue of the p-value). Second, an F_{dist} approach implemented in Arlequin V3.5 (Excoffier, Hofer, & Foll, 2009) was run using a hierarchical island model with 50,000 simulations, three simulated groups (i.e., three subspecies) and 100 demes per group. Finally, SNPs were identified as outliers when their F_{ST} value was inferior or superior to the 1st and 99th percentile of the F_{ST} simulated distribution, respectively. This outlier detection approach takes into account the population structure that generates an important excess of false-positive outliers when it is ignored (Excoffier et al., 2009). As neutral and outlier markers can reveal different genetic differentiation patterns (Luikart et al., 2003), analyses of population structure, population differentiation and population assignment were run on data sets with and without outlier SNPs.

2.7 | Population structure

Population structure was inferred using a model-based method employing a maximum-likelihood approach implemented in ADMIXTURE v1.3.0 (Alexander, Novembre, & Lange, 2009) and a k-means algorithmic method implemented in the *find.clusters* function of the *adegenet* R package (Jombart et al., 2010). ADMIXTURE uses the same statistical model as STRUCTURE (Falush, Stephens, & Pritchard, 2003), but runs faster as a result of a new optimized algorithm calculating ancestry. We ran ADMIXTURE with cross-validation for a number of groups (K) varying from 2 to 10. For each K value, calculations were repeated 10 times, using different random seeds to assess the stability of the estimate. The optimal K was identified as being the one exhibiting the lowest cross-validation error compared to other K values. The number of clusters was also assessed using a k-means method, a clustering algorithm which finds a given number K of groups maximizing the variation between groups. The *find.clusters* function (*adegenet* R package) was used to run sequentially the k-means algorithm with an increasing number of clusters K and to determine the optimal number of groups by the Bayesian information criterion method. The function was run several times to assess the stability of the optimal number of groups found. Before running the *find.clusters* function, missing values present in the data set were replaced by the mean allele frequency calculated for the entire set of individuals.

2.8 | Population differentiation and intrapopulation diversity

Population genetic diversity was evaluated by computing the observed (Ho; Nei, 1987) and expected (He; Nei, 1987) heterozygosity, and the inbreeding coefficient (Fis; Weir & Cockerham, 1984), using GenoDive 2.0b25 (Meirmans & Van Tienderen, 2004). Estimates of contemporary effective population size (Ne) were calculated for each population using the LD method (Waples & Do, 2008), as implemented in NeEstimator v2.01 (Do et al., 2014).

The extent of pairwise population differentiation was quantified through the computation of the unbiased F_{ST} estimator θ (Weir & Cockerham, 1984). Significance was determined by running 1,000 permutations using GenoDive 2.0b25 (Meirmans & Van Tienderen, 2004) and assessed against an FDR-adjusted p-value to account for multiple testing (Benjamini & Hochberg, 1994). A UPGMA dendrogram based on F_{ST} values was generated using the *hclust* function in the *stats* R package. A heatmap organized by subspecies and geographic origins was produced to illustrate population pairwise F_{ST}, and a hierarchical analysis of molecular variance (AMOVA) was computed among populations nested within subspecies groups (Excoffier, Smouse, & Quattro, 1992). Finally, a PCA was conducted on genotypes to summarize the overall variability among individuals. This PCA was computed using the *dudi.pca* function implemented in the ade4 R package (Dray & Dufour, 2007).

2.9 | TaqMan PCR assay vs. genotyping-by-sequencing results

Our team recently developed a qPCR-based suite of assays aimed at (among others things) distinguishing *L. d. dispar* specimens from those of *L. d. asiatica* and *L. d. japonica*, and assessing the presence of Asian introgression in material identified as *L. d. dispar* on the basis of a mitochondrial marker (Stewart et al., 2016). These assays are based on the presence of SNPs in the *cytochrome c oxidase I* (COI) gene as well as on the detection of North American "N" and Asian "A" alleles of the "FS1" nuclear marker (Garner & Slavicek, 1996). To generate an alternative genotypic characterization of the moths used here and to compare it with that obtained from genomewide GBS-derived SNPs, the TaqMan assays were run on all samples considered to be *L. d. dispar* as well as on those from the central Russian population (RB), considered to be *L. d. asiatica*, but found in the vicinity of the putative geographic boundary between *L. d. dispar* and *L. d. asiatica.* Assays were run as described in Stewart et al. (2016).

2.10 | Moth assignment to population

To perform assignment tests, we first employed discriminant analysis of principal components (DAPC; Jombart et al., 2010) which has the property of effectively highlighting genetic differentiation among groups while overlooking within-group variation (Jombart et al., 2010). DAPC is a multivariate approach wherein a discriminant analysis is conducted on the scores of a PCA computed on the raw SNP data. DAPC was first computed on 2,327 SNPs (both neutral and outlier SNPs identified previously) with eight prior groups corresponding to the populations studied here. DAPC computation was carried out using the function *dapc* implemented in the R package *adegenet* (Jombart et al., 2010). The pertinent number of principal components retained for DAPC analysis was submitted to a cross-validation test using the R function *xvalDapc* (R package *adegenet* Jombart et al., 2010). To rank SNPs according to their discriminant power, we relied on the SNP contribution for each of the seven DAPC axes. Thus, for each DAPC axis, SNP contributions were multiplied by the percentage of variation explained by the axis. SNPs were then ranked according to the value of the sum of their seven "weighted" contributions. Four panels, each comprising a different number of SNPs (12, 24, 48 and

96), were then developed from the top-ranked SNPs, plus a fifth panel comprising all SNPs.

As recommended by Anderson (2010), DAPC computation and ranking of SNPs were based on a *training set* of individuals (82% of individuals, i.e., 75), while the assignment power of the different SNP panels was assessed using the remaining 16 individuals constituting the *holdout set*. This Simple Training and Holdout method avoids upward assignment biases occurring when the same individuals are used to rank the SNPs and to test the assignment power of these SNPs (Anderson, 2010). The assignment of individuals from the *holdout set* to a given population was computed using the R function *predict.dapc* (R package *adegenet*; Jombart et al., 2010), which is based on the outcome of the DAPC analysis. To assess consistency of the assignment results, calculations were made for 10 different *training sets/holdout sets*, for which individuals were randomly sampled.

For comparative purposes, we also used a Bayesian assignment approach implemented in the *gsi_sim* Genetic Stock Identification software (Anderson et al., 2008), available in the *assigner* R package (function *assignment_ngs*; Gosselin, 2016). Assignment success was evaluated using the same 10 *training sets/holdout sets* used for the above DAPC analyses as well as panels of 12, 24, 48 and 96 top-ranking SNPs, the selection of which was here based on F_{ST} values. The entire SNP data set was also considered.

Beyond the development of a SNP panel for discriminating geographic populations, we carried out another DAPC analysis with the aim of identifying SNPs distinguishing European gypsy moth populations previously characterized to feature flightless females (UC and KG) from populations described as having flight-capable females, including one considered European gypsy moth (LJ) (Keena et al., 2008).

3 | RESULTS

3.1 | Genotyping and SNP filtering

We obtained an average of 84 million reads for the three sequencing runs. The UNEAK pipeline identified an average of 411,000 tags (unique sequences) per individual, computed from 2,370,000 reads. Of the 58,309 SNPs identified by UNEAK, 2,327 SNPs remained after applying the filtering procedure (Table 2). As expected for GBS, the filtering step that discarded the most SNPs was the one related to coverage (here, SNPs needed to be present in ≥80% of individuals and in at least seven populations). Interestingly, SNPs that were identified as deviating from the Hardy–Weinberg equilibrium were largely the result of heterogeneous read coverage among individuals; individuals with low coverage at a given locus may appear homozygous when, in fact, they are heterozygous, creating a departure from Hardy–Weinberg proportions. Linkage disequilibrium analysis identified 35 pairs of SNPs as highly linked, that is, with LD $r^2 > 0.80$ in at least three populations. Close examination of these SNP pairs revealed the presence, in about 50% of them, of an indel within a mononucleotide repetitive region located in an otherwise identical sequence background. When the LD filtering criterion was further constrained to $r^2 > 0.80$ in at least four populations, the 50% proportion increased to 95%. The UNEAK pipeline allows only one mismatch between sequences to call a SNP. Thus, when a sequencing error produces an indel in both alleles of a locus, the number of mismatch among alleles increases and, as a consequence, UNEAK considers the alleles with an indel as originating from a different locus (Data S1). For each SNP pair highly linked, the SNP presenting less missing data and supported by a greater number of reads was kept. At the end of the filtering process, five individuals (5%) presented a proportion of missing data >30% and were removed from the data set (Table 1).

3.2 | Linkage disequilibrium network analyses

We first explored variation in the topology of linkage disequilibrium networks and in the number of single-outlier clusters (SOCs) identified as a function of the values given to the φ and |E|min parameters. Network topology was stable at all parameter values tested, but the network generated with φ = 23 and |E|min = 3 was chosen here for illustrative purposes as it provided the best overall representation of the networks and SOCs obtained with the different combinations of parameter values (Figure 1a). Eight SOCs were identified containing between 12 and 178 SNPs (for details, see Data S2). Principal component analysis (PCA) of these SNPs indicated that all identified SOCs resulted from population structuring. For example, a PCA conducted on the SOC designated 315_0.65 (315 is the cluster number and 0.65 is the LD value where it merged with another cluster) indicated that the 43 SNPs found in this SOC distinguished the Russian population (RB) from the other populations (Figure 1b). PCAs carried out on each SOC revealed the same "one population vs. all other populations" pattern, excepted for SOC 222_0.74, where a PCA carried out on its 31 SNPs showed that the LJ population had allele frequencies intermediate between other *L. d. dispar* individuals and *L. d. asiatica/L. d. japonica* populations (Figure 1c). The PCA plots in Figure 1b revealed a clear differentiation of one individual (UC7M) relative to other moths in the population from which it was sampled (UC: Connecticut), where this "outsider" appeared closer to the *L. d. asiatica/L. d. japonica* populations. This trend was also observed in a PCA carried out on SOC 221_0.74, which singled out the Connecticut (UC) population (plot not shown).

3.3 | Identifying outlier SNPs

Of the 2,327 SNPs remaining after the filtering procedure (Table 2), 31 (1.3%) were identified by BayeScan 2.0 as having extreme F_{ST} values, while the F_{dist} approach implemented in Arlequin V3.5 identified 124 SNPs (5.2%) as outliers (Data S3). The proportion of SNPs identified as having extremely high F_{ST} values by Arlequin was higher (90/124) than the proportion computed by BayeScan (10/31). For the purpose of our study, we combined the results of both analyses and defined 133 SNPs as outliers. Subsequent analyses were carried out on both the 2,194 neutral SNP data set (94%) and the full (i.e., neutral + outlier) 2,327 SNP data set, as neutral and outlier SNPs can reveal different genetic differentiation patterns.

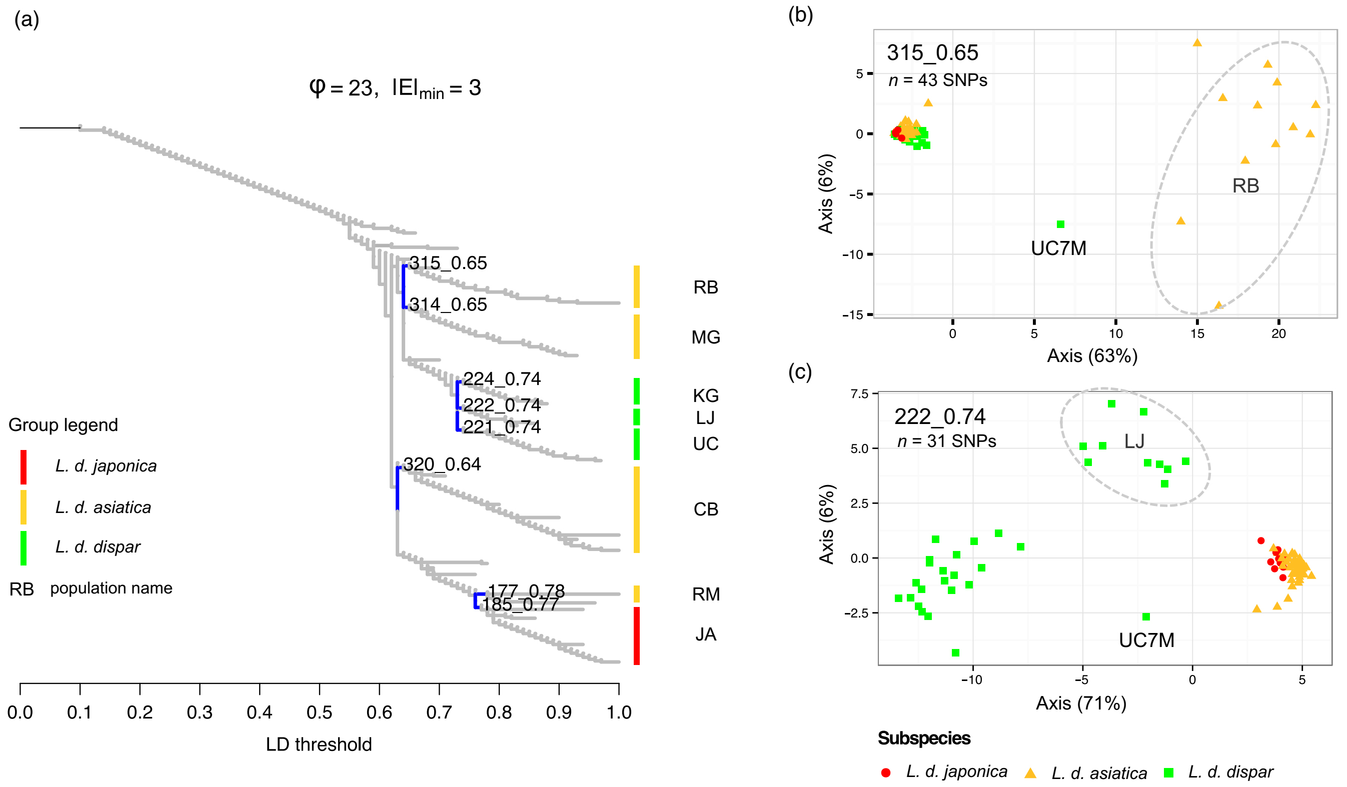

FIGURE 1 Linkage disequilibrium network analysis (LDna) applied to gypsy moth population genomic data. (a) Clustering tree of pairwise LD values from the 2,387 filtered SNPs from eight *L. dispar* populations (φ = 23, |E|min = 3). Branches corresponding to single-outlier clusters (SOCs) are highlighted in blue and labelled with their individual designation (e.g., in 315_0.65, 315 = cluster number and 0.65 = the LD threshold at which the SOC merged with another cluster). All identified SOCs result from population structuring and the population associated with each SOC is indicated on the right of the tree (see Table 1 for details on population names). (b, c) Two examples of principal component analysis (PCA) carried out on markers associated with SOCs identified in (a), the results of which led to the conclusion that population structuring is the evolutionary phenomenon that yielded the observed SOCs. For each PCA plot, data points belonging to the population associated with the featured SOC (RB and LJ in (b) and (c), respectively) are circled with a grey dashed line. One individual (UC7M) that showed a clear differentiation from other individuals in the source population (UC) is labelled separately

3.4 | Population structure

Analysis of the 2,194 neutral SNP data set using ADMIXTURE enabled the identification of eight populations; cross-validation error values for $K = 8$ were the lowest and the least variable (Figure 2a). The ADMIXTURE plots showed that the eight groups identified corresponded to the sampled populations (Figure 2b). In the Connecticut population (USA), one moth, UC7M, stood out as it displayed some genetic similarity with the central Russian population (RB), as noted above with LDna analysis. This population structure ($K = 8$) was generated nine times of 10 in replicated analytical runs; the remaining run, which displayed the highest cross-validation error, showed the Japanese population (JA) as being divided into two groups and the Lithuania one (LJ) composed of individuals with mixed ancestry from Greek (KG) and Asian populations (data not shown). At $K = 7$, where the cross-validation error was almost as low as for $K = 8$, but more variable (Figure 2a), the six runs with the lowest error generated a population structure where moths from the Lithuanian population (LJ) were grouped predominantly with those of the Greek population (KG), while showing some mixed ancestry with populations from Central Asia (RB and MG) and the USA (UC) (Figure 2b). At $K = 3$, where the parameter value corresponds to the number of subspecies, the three lineages could be well identified in the most frequent ADMIXTURE plot (7 runs of 10, cross-validation error <0.549), but with obvious introgression in populations near the geographic boundaries between subspecies (Figure 2c, bottom panel). For the remaining three runs at $K = 3$ (cross-validation error ≥0.549), the structures generated by the analysis did not show as good a match to the subspecies delimitation as the first one, particularly in the case of the Asian populations, which split into two groups, that is, Central Asia (RB and MG) and East Asia (RM and CB; Figure 2c, top panels). At $K = 4$, eight of the ten runs showed the Asian populations as divided into Central Asian and East Asian populations, while the *L. d. dispar* populations and the *L. d. japonica* populations each formed a separate group. However, the remaining two runs (cross-validation error ≥0.508) showed the North American population (UC) as distinct from the other populations, which were grouped according to their subspecies affiliation (see Data S4).

The genetic split between the eight sampled populations was also discerned using DAPC. Once again, a small number of the replicated runs pointed to different optimal K values (7 and 9). For $K = 7$, the populations from Greece (KG) and Lithuania (LJ) were grouped together as shown with ADMIXTURE (Figure 2b). For $K = 9$, the moth UC7M, identified by ADMIXTURE as having mixed ancestry, formed here a group by itself (data not shown).

We also ran ADMIXTURE and DAPC analyses on outlier SNPs only and on both neutral and outlier SNPs. In all cases, the number of populations detected was also $k = 8$, corresponding to the populations studied (Data S5.1).

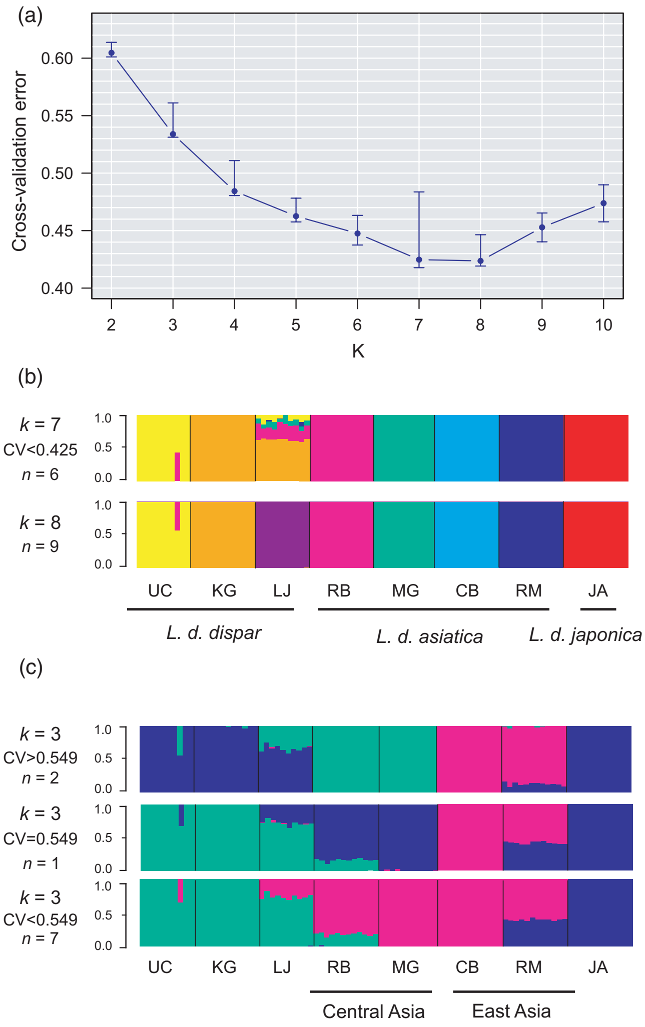

FIGURE 2 Gypsy moth population structure analysis using the ADMIXTURE software. Analysis conducted on 2,194 neutral SNPs derived from 91 *L. dispar* moths sampled in eight populations. (a) Cross-validation plot for *K* values from 2 to 10; for each *K* value, the dot represents the median of the cross-validation error calculated for 10 replicated computations while the vertical bars show the range of error values. (b) Membership coefficient plot for $K = 8$ (population number) and for $K = 7$ in a set of six runs in which the cross-validation error was equivalent to that assessed for $K = 8$. (c) Membership coefficient plots for $K = 3$ (subspecies number) as a function of the cross-validation error. For three runs (CV ≥ 0.549), the population structure did not match subspecies delimitations and the Asian populations were split into two geographic groups, that is, Central Asia (RB and MG) and East Asia (RM and CB). In (b) and (c), *n* is the number of occurrences of a given structure among the 10 replicated computations

3.5 | Population differentiation and intrapopulation diversity

Fixation indices (F_{ST}) among all pairwise populations were significant ($p < .001$; Figure 3). Mean F_{ST} across all 2,194 neutral SNPs was 0.420, and pairwise comparisons among the eight sampled populations ranged from 0.295 (KG vs. LJ) to 0.567 (UC vs. JA) (Figure 3; Data S6). Both the heatmap and the F_{ST}-based dendrogram separated populations according to their subspecies designations. F_{ST} values among populations within subspecies were lower than those measured between populations belonging to different subspecies. The Lithuanian (LJ) and far east Russian (RM) populations, originally sampled close to subspecies geographic boundaries, revealed overall lower F_{ST}. The AMOVA showed significant genetic differentiation among subspecies ($F_{ST} = 0.131$, *p*-value = .004) as well as among populations within each subspecies ($F_{ST} = 0.311$, *p*-value = .001). PCA performed on genotypes (neutral SNPs) also revealed clustering of populations according to their putative subspecies affiliation; the first two main principal component axes explained 28.37% of the total genetic differentiation (Figure 4). More specifically, the first axis separated populations of *L. d. dispar* from those of *L. d. asiatica* and *L. d. japonica*, whereas the second axis enabled discrimination between the *L. d. japonica* population and *L. d. asiatica*. Among *L. d. dispar* moths, those from Lithuania (LJ) showed the greatest genetic proximity to moths from *L. d. asiatica* populations. Conversely, the North American population (UC) was seen to be the most differentiated from the four *L. d. asiatica* populations. Coherently

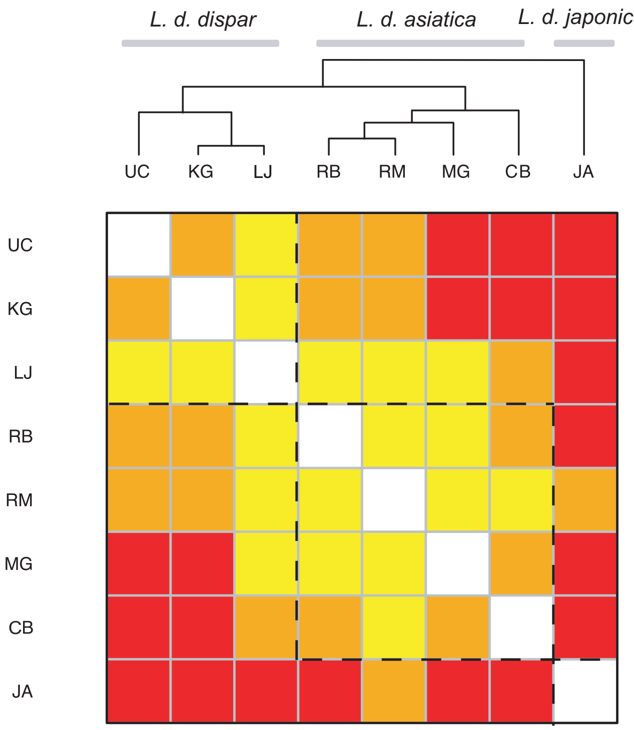

FIGURE 3 Gypsy moth F_{ST} population analysis. F_{ST} population dendrogram and heatmap of F_{ST} pairwise values among eight *L. dispar* populations. Heatmap colour code: yellow, $F_{ST} = 0–0.39$; orange, $F_{ST} = 0.39–0.45$; red, $F_{ST} = 0.45–0.57$. Dashed black lines delimit intra-subspecies pairwise F_{ST} values

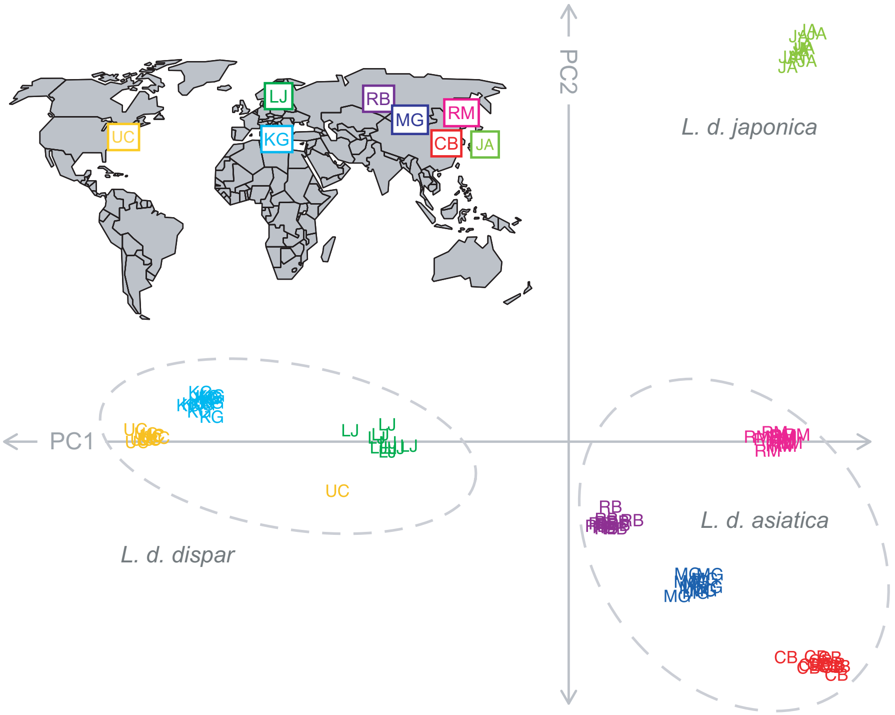

FIGURE 4 Upper left: sampling locations of *Lymantria dispar* moths used in this study. Refer to Table 1 for details on names of each location (boxes). Main plot: principal component analysis (PCA) applied to 2,194 neutral SNPs reveals distinct coordinates for each population and clustering of populations according to their putative subspecies affiliation. The first and second principal component axes explained 15.06% and 12.89% of the total genetic differentiation, respectively

with the LDna and structure analyses, the UC7M individual displayed some genetic similarity to moths from the central Russian population (RB). At this stage, we cannot dismiss the possibility of a procedural error that led to the inclusion, in our UC sample, of a moth from a different source population. Among *L. d. asiatica* populations, those from Mongolia and China (MG and CB) displayed distinct separation from the two Russian populations (RM and RB), with RB being the least differentiated from *L. d. dispar*. A PCA computed on both neutral and outlier SNPs generated an identical pattern of population differentiation (Data S5.2). Interestingly, a PCA performed just on outlier SNPs also distinguished subspecies, but the two-first axes could not discriminate some populations, that is, the Connecticut (UC) from the Greek populations in the *L. d. dispar* subspecies and the Chinese, Siberian and Mongolian populations in the *L. d. asiatica* subspecies. The Russian (RB and RM) and Lithuanian (LJ) populations showed the highest degree of genetic diversity (He) while the population from Connecticut (UC) displayed the lowest (Table 1). All populations showed significant positive Fis values, revealing heterozygote deficiency, but the extent of the deficit varied greatly, from 0.095 in the Mongolian population to a three-fold higher value (0.306) in the Lithuanian population. Assessments of effective size (Ne) also displayed very significant variation, with values ranging from 2.8 to 139.5 for the Connecticut (UC) and Siberian (RB) populations, respectively. No correlation seemed to link the Fis and Ne values with the number of egg masses sampled to initiate the laboratory colonies or the number of generations bred in the laboratory.

3.6 | Impact of the absence of LD filtering

The presence of high-LD SNPs did not significantly affect the results of our outlier SNP detection procedure. When Arlequin and BayeScan were run several times on data sets with and without high-LD SNPs, the SNPs showing the highest F_{ST} values were the same. In addition, variation in the list of outlier SNPs among runs was similar for both data sets and each outlier detection method. Thus, for further assessments of the impact of the presence of high-LD SNPs on analytical results, we deleted from the high-LD SNP data set the same outlier SNPs identified from the data set without high-LD SNPs (see above Identifying outlier SNPs section). Characterization of population structure using ADMIXTURE was not affected by the presence high-LD SNPs and, as expected, other analyses not requiring independence of SNPs (i.e., LDna, F_{ST} calculations, genetic diversity indices, DAPC) generated similar results with both data sets.

3.7 | TaqMan PCR assay vs. genotyping-by-sequencing (GBS) results

Using the COI-based TaqMan assay of Stewart et al. (2016), moths from Connecticut (UC), Greece (KG) and Lithuania (LJ) were assigned to the *L. dispar dispar* subspecies (Table 3), in line with results of analyses based on genomewide SNPs (e.g., Figures 2 and 4). However, the same assay identified the central Russian population (RB) as *L. dispar dispar* whereas analysis of GBS-derived SNPs strongly suggested it belonged to the *L. dispar asiatica* subspecies (see Figures 2 and 4).

For an independent assessment of the occurrence of Asian introgression into EGM, we used the FS1 nuclear marker assay of Stewart et al. (2016), which detects the presence of North American "N" and Asian "A" FS1 alleles in unknown samples. The A allele was detected in all four populations examined (Table 3) and its frequency increased from west to east (20%–100%). In the Greek (KG) and Lithuanian (LJ) populations, the A allele was the major allele (55% and 87.5%,

TABLE 3 FS1- and COI-based genotypes of the three *L. d. dispar* and the RB *L. d. asiatica* populations used in this study, arranged by longitude from east to west

		FS1 Genotypes[a]				FS1 Allelic frequency		COI gene assay	
Population	N[a]	AA	AN	NN	Null	A	N	*L. d. dispar*	*L. d. asiatica*
RB	12 (12)	8	0	0	4	100	0	12	0
LJ	12 (12)	9	3	0	0	87.5	12.5	12	0
KG	12 (12)	2	9	1	0	55	45	12	0
UC	12 (12)	0	4	8	0	20	80	12	0

[a] No. of moths from which DNA was extracted (No. of samples successfully genotyped).

respectively), flagging these two European populations as displaying significant Asian introgression. By contrast, analysis of genomewide SNPs did not detect Asian introgression in the Connecticut (UC) and Greek (KG) populations, while the Lithuanian (LJ) displayed moderate Asian introgression, but less than what the FS1 genotype might suggest. The UC7M individual found to display Asian admixture using GBS-derived SNPs was heterozygous (A/N) for the FS1 marker, as were three other specimens from the Connecticut population. Earlier FS1 genotyping of this population revealed a 20% proportion of A/N genotype (Keena et al., 2008).

3.8 | Moth assignment to population

With the exception of results obtained for the 12-SNP panel, the assignment success of individuals to their respective population was high (≥86.25%), irrespective of the number of SNPs and method used (Figure 5a); however, the mean assignment success to population was low (51.25%) for the 12-SNP panel, following computation using the gsi_sim method. An assignment success of 100% was obtained for the 48-SNP panel using the gsi_sim method while the same level of success required the full SNP data set when using DAPC. However, the DAPC method outperformed gsi_sim for panels containing fewer than 48 SNPs (Figure 5a).

We conducted a DAPC using priors corresponding to populations characterized earlier as having either flightless or flight-capable females. This analysis yielded two SNPs whose allele frequency correlated with the presence of flying females in these populations (Figure 5b). Only one individual in each of the UC and LJ populations had a discordant haplotype for TP59129. For TP7787, one or two discordant haplotypes were observed for the LJ and RB populations, respectively. A blastx search against the NCBI nonredundant database, using the reads bearing these SNPs as query, failed to produce significant matches to known proteins.

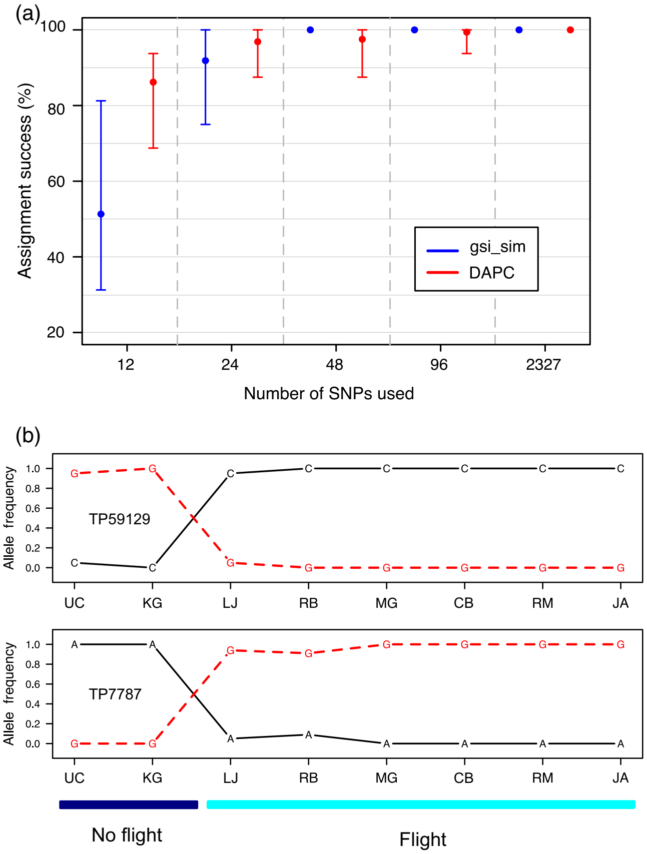

FIGURE 5 Assessment of the accuracy of panels of SNPs in identifying the geographic origins of gypsy moth samples. (a) Assignment success as a function of increasing size of SNP panel, as assessed using two different methods: a multivariate approach based on discriminant analysis of principal component (DAPC) and a Bayesian approach implemented in the Genetic Stock Identification program gsi_sim. SNPs constituting the panels were selected according to their contribution to the discriminant axes, for the DAPC approach, and according to their F_{ST} values for the gsi-sim program (see Materials and methods section for details). Each dot represents the mean assignment success, using the Simple Training and Holdout (STH) method, computed for 10 replicated SNP panels; each vertical bar shows the range of values of the mean assignment success obtained. (b) Allele frequency of two SNPs, designated TP59129 and TP7787, proposed here as candidates for the identification of gypsy moth populations known to have either flightless or flight-capable females

4 | DISCUSSION

The aim of the present proof-of-concept study was to assess the feasibility of using GBS-derived SNPs to identify source populations of gypsy moth samples. The success of such a strategy is heavily dependent on the degree of genetic differentiation among the biological entities being considered (Cornuet, Piry, Luikart, Estoup, & Solignac, 1999). For example, for populations that are relatively well differentiated (e.g., F_{ST} = 0.11), such as European cattle breeds, the correct assignment of an individual to its source breed can reach 85% with only 90 SNPs (Negrini et al., 2009). By contrast, for populations of the American lobster, which display minimal genetic differentiation (F_{ST} = 0.002), assignment success has been shown to plateau at 31%, using as many as 10,156 SNPs (Benestan et al., 2016). The different analyses conducted in the present study revealed strong differentiation among the eight gypsy moth populations examined (minimum F_{ST} value = 0.237). As a consequence, we obtained 100% assignment success using all 2,327 SNPs (neutral and outlier), irrespective of the assignment method employed (i.e., DAPC or gsi_sim). In addition, assignment success remained generally high (86%–100%) using fewer SNPs (12, 24, 48 and 96), although the Bayesian method generated a lower score (51%) for the 12-SNP panel (Figure 5a). The two assignment methods tested here use different approaches for generating SNP panels. Whereas DAPC ranks SNPs according to their discriminant power, gsi_sim ranks them according to their F_{ST} values; panels are then developed beginning with the most discriminating SNPs and those displaying the highest F_{ST} values, respectively. As a result, the 12-SNP panels generated by gsi_sim contained only high F_{ST} outlier SNPs, which were here shown to be less powerful than neutral SNPs in discriminating some populations (Data S5.3). This would explain the better performance of DAPC in tests conducted on the smallest SNP panel, which contained both neutral and high F_{ST} outlier SNPs.

Given that the present study was based on populations reared in the laboratory for several generations, the question arises as to how different our assessment of assignment success, as well as the genetic structure and diversity we obtained, would have been had we applied the procedure to wild populations. The high F_{ST} values we observed among our experimental populations are expected to have been driven by different forces, including the large geographic spread between the source populations. However, this high degree of genetic differentiation, necessary for high assignment success, could have been enhanced by accelerated changes in allele frequency caused by a founder effect and by the genetic drift that affected our small populations reared in the laboratory over several years (Allendorf, Luikart, & Aitken, 2012). A mean pairwise F_{ST} value of 0.284 was reported by Keena et al. (2008) for six of the populations used in the present study (see Table 1), a value computed one generation after initial field collection (see Table 5 in Keena et al. (2008)). Although comparison of F_{ST} values obtained using different types and numbers of markers is delicate, the above F_{ST} assessment is clearly lower than the one we computed for the same six populations after 20 years of laboratory rearing (0.388; Data S6). Thus, despite the extensive measures taken to limit the impact of laboratory rearing on the populations we used (see Materials and methods for details), our test populations may have displayed more pronounced genetic differentiation than those of the original wild stocks at the time they were collected. It follows that the assignment success reported here would likely be lower if the F_{ST} assessment were to be repeated using SNPs derived from wild populations, particularly in considering the smallest SNP panels. However, previous work revealed important genetic differentiation among natural gypsy moth populations, with mean F_{ST} values of 0.210 and 0.192 reported by Keena et al. (2008) and Wu et al. (2015), respectively. It is worth pointing out that with twice-lower F_{ST} values, the correct assignment of a bovine to its source breed reached 85% using 90 SNPs (Negrini et al., 2009). Thus, the high degree of genetic differentiation observed in natural gypsy moth populations suggests that an assignment procedure based on GBS-derived SNPs applied to intercepted moths has a high chance of success.

With respect to gypsy moth population structure, our study supports conclusions made earlier by other workers (Keena et al., 2008; Wu et al., 2015), including a clear delineation of the three subspecies (F_{ST} values and structure analysis based on 2,194 neutral SNPs for K = 3; Figures 2, 3 and 4), with apparent introgression at the geographic boundaries between subspecies (here, in the populations from Lithuania (LJ) and central Russia (RB), for the *L. d. dispar*/*L. d. asiatica* boundary, and in the population from the Russian Far East (RM) for the *L. d. asiatica*/*L. d. japonica* boundary). Higher levels of introgression in these populations are also supported by higher genetic diversity indices (He; Table 1), as expected for populations in hybrid zones. Thus, 20 years of laboratory rearing does not appear to have significantly altered the main patterns of gypsy moth population structure. One exception, however, is the status of the North American population, which differs between our study and the two above-cited reports. Based on nine microsatellites, Wu et al. (2015) identified North American gypsy moths as a distinct genetic entity, that is, in addition to the three recognized subspecies and their hybrids. Contrastingly, our genomewide SNP-based analyses did not clearly distinguish North American moths from European *L. d. dispar* populations (see Figures 2, 3 and 4, and Data S4). Keena et al. (2008) assessed the genetic diversity of worldwide gypsy moth populations, including the North American source used here (UC; genetic diversity estimates calculated on 1st/2nd generation after establishment of laboratory rearing), using six markers, four of which were also used by Wu et al. (2015). Interestingly, the status of the North American population varied as a function of the markers used in the analyses. For example, when a mitochondrial marker was included, the North American population was deemed distinct from European *L. dispar* populations, whereas it formed a single group with the European moths when only nuclear markers were considered (Keena et al., 2008). It is difficult to say whether this outcome is due to cyto-nuclear discordance as only one mitochondrial marker was used in this particular instance. However, this comparison illustrates the fact that when only a few markers are considered, some patterns of population structure can be overemphasized due to a very low proportion of the genome being sampled. Future work based on genomewide SNPs and a larger sample of contemporary North American and European specimens should help clarify the extent to which the North American population is genetically distinct from its European progenitor.

Sample size (10–12 moths/population) is another feature of our study design that had the potential of biasing our evaluation of population genetic differentiation and genetic diversity. Although they could be considered relatively small, our sample sizes were initially deemed sufficient to achieve our objectives given that the large number of SNPs generated by high-throughput sequencing tends to lessen the requirement for large sample sizes (Jeffries et al., 2016). For example, a simulation study revealed that accurate estimates of population differentiation could be obtained from small sample sizes (n = 4–6) if a large number of markers were considered (>1,000 SNPs; Willing et al., 2012). An empirical study on an Amazonian plant species provided support for this assessment, generating accurate estimates of F_{ST} using ≥1,500 SNPs and ≥8 individuals per population (Nazareno et al., 2017). The results of these studies suggest that the genetic diversity estimates reported here (F_{ST}, F_{IS}, etc.) are reliable as they are based on 10–12 individuals/population and 2,327 SNPs.

For comparative purposes, we assessed the nuclear FS1 genotype of all *L. d. dispar* moths included in our study, using the TaqMan assay of Stewart et al. (2016), to assess the occurrence of Asian introgression into these populations, considered here to be European gypsy moths on the basis of their mitochondrial COI haplotype (Table 3). The presence of the Asian ("A") FS1 allele in a significant proportion of our three *L. d. dispar* populations (UC, KG, LJ) suggested a degree of Asian introgression perhaps greater than that assessed using the SNP data. For example, 90% of the moths from the Greek population (KG) had an FS1 A allele, whereas analyses based on SNP data revealed no Asian introgression in these insects (Figure 2). These contrasting results indicate that the FS1 nuclear marker may generate overestimates of Asian introgression into *L. d. dispar*. This observation should be taken into account in making decisions about EGM material flagged as having an FS1 A allele, so as to avoid unjustified entry refusal in Canada or the United States of vessels found to be infested with such moths.

Although female flight is believed to be a trait observed uniquely in Asian gypsy moth subspecies, studies on the world distribution of this trait have revealed that *L. d. dispar* populations from the north-eastern parts of Europe possess gliding and flight-capable females (Keena et al., 2008; Reineke & Zebitz, 1998). This observation raises regulatory concern as 20% of merchandize imported on European vessels originates from north-eastern Europe, and currently vessels from this part of Europe are not subject to phytosanitary regulation for gypsy moth (CFIA, 2014). In our data set, we identified two high F_{ST} outlier markers that enable separation of moths of the North American and Western Europe populations with flightless female (UC, KG) from those of north-eastern Europe (LJ) and Asian populations with flight-capable female. The sequences containing these SNPs did not reveal significant matches to known proteins, so future work will aim to determine whether these two SNPs are present in genes or genomic regions linked to female flight capacity or reveal population structure due to other selected traits. Although the usefulness of these two markers as predictors of female flight capability appears limited, their identification here suggests that research aimed at localizing the genomic determinants of flight capacity using other genomewide SNP-based approaches (e.g., QTL, GWAS) will likely be rewarding.

High linkage disequilibrium (LD) between markers can introduce bias in analyses requiring independence of loci, such as outlier SNP detection or model-based methods to characterize population structure. For example, LD between markers in close proximity in a genome ("background LD" or BLD) could cause overestimation of population divergence and incorrect estimation of population admixture using the STRUCTURE software (Falush et al., 2003). Given the uncertainty about the type of bias caused by SNPs in high LD in genomic analyses (Pérez-Figueroa et al., 2010), the presence of some high-LD SNPs in our data set provided an opportunity to assess the influence of these SNPs on the outcome of our outlier identification and population structure analyses. Interestingly, the presence of high-LD SNPs failed to cause significant variation in either the outlier SNP identification results or in population structure characterization. This absence of effect may be due to the low proportion of high-LD SNPs in our data set (1.5%) and their assumed random distribution along the genome. Indeed, in such situations, model-based methods used to evaluate population structure seem to perform reasonably well, irrespective of the presence of some high-LD SNPs in the data set (Falush et al., 2003).

In conclusion, the present proof-of-concept study demonstrates the power of a genomewide SNP-based approach to correctly assign gypsy moths to their source populations. Our work also shows that when populations are well differentiated (F_{ST} ≥ 0.237), high assignment success (>81.88%) can be achieved using as few as 24 SNPs. However, because our analyses are based on laboratory-reared populations, the SNP panels we developed are not immediately applicable to wild gypsy moth populations. We are now setting out to repeat the present work using a large panel (>50) of contemporary, natural populations, with numerous individuals per population (30–40), which will enable an accurate assessment of population allele frequencies. Because natural gypsy moth populations are expected to be less well differentiated than the populations examined here, we will likely need more SNPs for successful discrimination of populations. In this respect, the foreseeable availability of a gypsy moth reference genome to map GBS reads should greatly increase the number of SNPs available for such discrimination. In addition, recently commercialized tools for target enrichment (e.g., AmpliSeq™) are expected to provide a high degree of latitude with respect to the number of SNPs that can be selected for rapid sequencing in a single run. Altogether, the above considerations offer great hope for the development of a molecular tool capable of accurately assigning intercepted gypsy moths to their source populations.

ACKNOWLEDGEMENTS

We thank H. Nadel, M. Reed (USDA-APHIS) and P. Moore (USDA Forest Service) for providing the moths used in this study. This research was supported by grants from the Genomics Research and Development Initiative (GRDI; Government of Canada), and the Natural Sciences and Engineering Research Council of Canada (NSERC; Discovery Grant 171350-2012) to Michel Cusson. This work was also supported by grants from Genome Canada, Genome British Columbia, and Génome Québec (Genomic Applications Partnership Program, Grant #244DIA and Large Scale Applied Research Program, Grant #182IAS).

DATA ARCHIVING STATEMENT

Data for this study are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.01s00.

ORCID

Sandrine Picq http://orcid.org/0000-0003-1436-3307
Melody Keena http://orcid.org/0000-0003-3099-6243
Michel Cusson http://orcid.org/0000-0003-1541-6052
Nathan Havill http://orcid.org/0000-0002-4004-8266

REFERENCES

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

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2012). *Conservation and the genetics of populations* (2nd edn). Oxford: Blackwell Publishing Ltd.

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: Standard methods are upwardly biased. *Molecular Ecology Resources*, *10*(4), 701–710. https://doi.org/10.1111/j.1755-0998.2010.02846.x

Anderson, E. C., Waples, R. S., & Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1475–1486. https://doi.org/10.1139/F08-049

Andrews, S. (2016). FastQC: A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

APHIS-USDA (2006). Asian gypsy moth cooperative eradication program orange county, California. Retrieved from https://www.aphis.usda.gov/plant_health/ea/downloads/gypsymothorangecountyfinal.pdf

APHIS-USDA (2016). Asian gypsy moth (Pest Alert No. APHIS 81-35-027).

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190.

Aukema, J. E., Leung, B., Kovacs, K., Chivers, C., Britton, K. O., Englin, J., … Von Holle, B. (2011). Economic impacts of non-native forest insects in the continental United States. *PLoS ONE*, *6*(9), e24587. https://doi.org/10.1371/journal.pone.0024587

Benestan, L. M., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2016). Erratum Benestan et al. 2015. *Molecular Ecology*, *25*(7), 1626–1629. https://doi.org/10.1111/mec.13600

Benjamini, Y., & Hochberg, Y. (1994). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. http://www.jstor.org/stable/2346101

Bigsby, K. M., Tobin, P. C., & Sills, E. O. (2011). Anthropogenic drivers of gypsy moth spread. *Biological Invasions*, *13*(9), 2077. https://doi.org/10.1007/s10530-011-0027-6

Bogdanowicz, S. M., Schaefer, P. W., & Harrison, R. G. (2000). Mitochondrial DNA variation among worldwide populations of gypsy moths, *Lymantria dispar*. *Molecular Phylogenetics and Evolution*, *15*(3), 487–495. https://doi.org/10.1006/mpev.1999.0744

Bradshaw, C. J. A., Leroy, B., Bellard, C., Roiz, D., Albert, C., Fournier, A., … Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, *7*, 12986. Retrieved from https://doi.org/10.1038/ncomms12986

CFIA (2014). Canada-United States Joint AGM Industry Notice 2014.

Cornuet, J. M., Piry, S., Luikart, G., Estoup, A., & Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, *153*(4), 1989–2000.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

deWaard, J. R., Mitchell, A., Keena, M. A., Gopurenko, D., Boykin, L. M., Armstrong, K. F., … Humble, L. M. (2010). Towards a global barcode library for Lymantria (Lepidoptera: Lymantriinae) tussock moths of biosecurity concern. *PLoS ONE*, *5*(12), e14280. https://doi.org/10.1371/journal.pone.0014280

Do, C., Waples, R., Peel, D., Macbeth, G., Tillett, B., & Ovenden, J. (2014). NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources*, *14*(1), 209–214. https://doi.org/10.1111/1755-0998.12157

Dray, S., & Dufour, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *1*(4), 1–20. https://doi.org/10.18637/jss.v022.i04

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, *103*(4), 285–298. https://doi.org/10.1038/hdy.2009.74

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491. http://www.genetics.org/content/131/2/479

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462648/

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Garner, K. J., & Slavicek, J. M. (1996). Identification and characterization of a RAPD-PCR marker for distinguishing Asian and North American gypsy moths. *Insect Molecular Biology*, *5*(2), 81–91.

Gosselin, T. (2016). *Assigner: Assignment analysis with GBS/RAD data using R*. R package version 0.1.9. https://github.com/thierrygosselin/assigner. https://doi.org/10.5281/zenodo.46723

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, *11*, 117–122. https://doi.org/10.1111/j.1755-0998.2010.02967.x

Jeffries, D. L., Copp, G. H., Handley, L. L., Olsén, K. H., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius* L. *Molecular Ecology*, *25*(13), 2997–3018. https://doi.org/10.1111/mec.13613

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), 1–15. https://doi.org/10.1186/1471-2156-11-94

Keena, M. A., Côté, M.-J., Grinberg, P. S., & Wallner, W. E. (2008). World distribution of female flight and genetic variation in *Lymantria dispar* (Lepidoptera: Lymantriidae). *Environmental Entomology*, *37*(3), 636–649. https://doi.org/10.1603/0046-225X(2008)37[636:WDOFFA]2.0.CO;2

Keena, M. A., Grinberg, P. S., & Wallner, W. E. (2007). Inheritance of female flight in *Lymantria dispar* (Lepidoptera: Lymantriidae). *Environmental Entomology*, *36*(2), 484–494. https://doi.org//10.1093/ee/36.2.484

Kemppainen, P., Knight, C. G., Sarma, D. K., Hlaing, T., Prakash, A., Maung Maung, Y. N., … Walton, C. (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, *15*(5), 1031–1045. https://doi.org/10.1111/1755-0998.12369

Kenis, M., Auger-Rozenberg, M.-A., Roques, A., Timms, L., Péré, C., Cock, M. J. W., … Lopez-Vaamonde, C. (2009). Ecological effects of invasive alien insects. *Biological Invasions*, *11*(1), 21–45. https://doi.org/10.1007/s10530-008-9318-y

Lance, D., & Barbosa, P. (1982). Host tree influences on the dispersal of late instar gypsy moths, *Lymantria dispar*. *Oikos*, *38*(1), 1–7.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., & Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). *Evolutionary Application*, *7*(3), 355–369. https://doi.org/10.1111/eva.12128

Liebhold, A. M., Gottschalk, K. W., Muzika, R.-M., Montgomery, M. E., Young, R., O'Day, K., & Kelley, B. (1995). Suitability of North American tree species to the gypsy moth: a summary of field and laboratory tests. USDA Forest Service General Technical Report NE, 211.

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., … Costich, D. E. (2013). Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, *9*(1), e1003215. https://doi.org/10.1371/journal.pgen.1003215

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Mascher, M., Wu, S., Amand, P. S., Stein, N., & Poland, J. (2013). Application of genotyping-by-sequencing on semiconductor sequencing platforms: A comparison of genetic and reference-based marker ordering in barley. *PLoS ONE*, *8*(10), e76925. https://doi.org/10.1371/journal.pone.0076925

Meirmans, P. G., & Van Tienderen, P. H. (2004). genotype and genodive: Two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, *4*(4), 792–794. https://doi.org/10.1111/j.1471-8286.2004.00770.x

Morin, R. S., & Liebhold, A. M. (2016). Invasive forest defoliator contributes to the impending downward trend of oak dominance in eastern North America. *Forestry*, *89*(3), 284–289. https://doi.org/10.1093/forestry/cpv053

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. C. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Molecular Ecology Resources*. [Epub ahead of print]. https://doi.org/10.1111/1755-0998.12654

Negrini, R., Nicoloso, L., Crepaldi, P., Milanesi, E., Colli, L., Chegdani, F., … Ajmone Marsan, P. (2009). Assessing SNP markers for assigning individuals to cattle populations. *Animal Genetics*, *40*(1), 18–26. https://doi.org/10.1111/j.1365-2052.2008.01800.x

Nei, M. (1987). *Molecular evolutionary genetics*. New York, NY: Columbia University Press.

Nickason, M. (2001). *Dispersal patterns of gypsy moth larvae (Lymantria dispar) (Undergraduate Ecology Research Reports)* (p. 13). Corpus Christ, TX: Institute of Ecosystem Studies, Texas A&M University-Corpus Christ.

Pérez-Figueroa, A., García-Pereira, M. J., Saura, M., Rolán-Alvarez, E., & Caballero, A. (2010). Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, *23*(10), 2267–2276. https://doi.org/10.1111/j.1420-9101.2010.02093.x

Pfeifer, T. A., Humble, L. M., Ring, M., & Grigliatti, T. A. (1995). Characterization of gypsy moth populations and related species using a nuclear DNA marker. *The Canadian Entomologist*, *127*(01), 49–58.

Picq, S., McMillan, W. O., & Puebla, O. (2016). Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecology and Evolution*, *6*(7), 2109–2124. https://doi.org/10.1002/ece3.2028

Pogue, M. G., & Schaefer, P. W. (2007). *A review of selected species of Lymantria Hübner [1819] including three new species (Lepidoptera: Noctuidae: Lymantriinae) from subtropical and temperate regions of Asia, some potentially invasive to North America*. Morgantown, WV: United States Department of Agriculture Forest Service, Forest Health Technology Enterprise Team.

Puckett, E. E., & Eggert, L. S. (2016). Comparison of SNP and microsatellite genotyping panels for spatial assignment of individuals to natal range: A case study using the American black bear (*Ursus americanus*). *Biological Conservation*, *193*, 86–993. https://doi.org/10.1016/j.biocon.2015.11.020

Reineke, A., & Zebitz, C. P. W. (1998). Flight ability of gypsy moth females (*Lymantria dispar* L.) (Lep., Lymantriidae): A behavioural feature characterizing moths from Asia? *Journal of Applied Entomology*, *122*(1–5), 307–310. https://doi.org/10.1111/j.1439-0418.1998.tb01502.x

Roesti, M., Salzburger, W., & Berner, D. (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evolutionary Biology*, *12*(1), 1–7. https://doi.org/10.1186/1471-2148-12-94

Stewart, D., Zahiri, R., Djoumad, A., Freschi, L., Lamarche, J., Holden, D., … Cusson, M. (2016). A multi-species TaqMan PCR assay for the identification of Asian gypsy moths (*Lymantria* spp.) and other invasive lymantriines of biosecurity concern to North America. *PLoS ONE*, *11*(8), e0160878. https://doi.org/10.1371/journal.pone.0160878

Streifel, M. A. (2016). Invasion biology of the gypsy moth (*Lymantria dispar* (L.)) at a northern range boundary in Minnesota. M. S. thesis, University of Minnesota, St. Paul, USA.

Teo, Y. Y., Fry, A. E., Clark, T. G., Tai, E. S., & Seielstad, M. (2007). On the usage of HWE for identifying genotyping errors. *Annals of Human Genetics*, *71*(5), 701–703. https://doi.org/10.1111/j.1469-1809.2007.00356.x

Tobin, P. C., Bai, B. B., Eggen, D. A., & Leonard, D. S. (2012). The ecology, geopolitics, and economics of managing *Lymantria dispar* in the United States. *International Journal of Pest Management*, *58*(3), 195–210.

Tobin, P. C., Liebhold, A. M., & Anderson Roberts, E. (2007). Comparison of methods for estimating the spread of a non-indigenous species. *Journal of Biogeography*, *34*(2), 305–312. https://doi.org/10.1111/j.1365-2699.2006.01600.x

Waples, R. S., & Do, C. (2008). LDNE: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, *8*(4), 753–756. https://doi.org/10.1111/j.1755-0998.2007.02061.x

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Weseloh, R. M. (1997). Evidence for limited dispersal of larval gypsy moth, *Lymantria dispar* L. (Lepidoptera: Lymantriidae). *The Canadian Entomologist*, *129*(2), 355–361.

Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE*, *7*(8), e42649. https://doi.org/10.1371/journal.pone.0042649

Wu, Y., Molongoski, J. J., Winograd, D. F., Bogdanowicz, S. M., Louyakis, A. S., Lance, D. R., … Harrison, R. G. (2015). Genetic structure, admixture and invasion success in a Holarctic defoliator, the gypsy moth (*Lymantria dispar*, Lepidoptera: Erebidae). *Molecular Ecology*, *24*(6), 1275–1291.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Picq S, Keena M, Havill N, et al. Assessing the potential of genotyping-by-sequencing-derived single nucleotide polymorphisms to identify the geographic origins of intercepted gypsy moth (*Lymantria dispar*) specimens: A proof-of-concept study. *Evol Appl*. 2018;11:325–339. https://doi.org/10.1111/eva.12559