Deciphering the demographic history of allochronic differentiation in the pine processionary moth Thaumetopoea pityocampa

R. Leblois1,2* | M. Gautier1,2* | A. Rohfritsch1 | J. Foucaud1 | C. Burban3 | M. Galan1 | A. Loiseau1 | L. Sauné1 | M. Branco4 | K. Gharbi5 | R. Vitalis1,2 | C. Kerdelhuet1

1CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montferrier sur Lez Cedex, France
2Institut de Biologie Computationnelle (IBC), Université de Montpellier, Montpellier, France
3INRA, UMR1202 BIOGECO (INRA – Université de Bordeaux), Cestas Cedex, France
4Centro de Estudos Florestais (CEF), Instituto Superior de Agronomia (ISA), University of Lisbon, Lisbon, Portugal
5Edinburgh Genomics, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

Correspondence
Carole Kerdelhuet, CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montferrier sur Lez Cedex, France. Email: Carole.Kerdelhuet@inra.fr

Funding information
UK Medical Research Council, Grant/Award Number: G0900740; Agence Nationale de la Recherche, Grant/Award Number: GenoPheno Grant ANR-10-JCJC- 1705-01; Faculdade de Ciencias e Tecnologia, Universidade Nova de Lisboa, Grant/Award Number: UID/AGR/00239/2013; Institut National de la Recherche Agronomique, Grant/Award Number: GAPP project, INRA AIP BioRessources 2012; UK Natural Environment Research Council, Grant/Award Number: R8/H10/56

Abstract
Understanding the processes of adaptive divergence, which may ultimately lead to speciation, is a major question in evolutionary biology. Allochronic differentiation refers to a particular situation where gene flow is primarily impeded by temporal isolation between early and late reproducers. This process has been suggested to occur in a large array of organisms, even though it is still overlooked in the literature. We here focused on a well-documented case of incipient allochronic speciation in the winter pine processionary moth Thaumetopoea pityocampa. This species typically reproduces in summer and larval development occurs throughout autumn and winter. A unique, phenologically shifted population (SP) was discovered in 1997 in Portugal. It was proved to be strongly differentiated from the sympatric "winter population" (WP), but its evolutionary history could only now be explored. We took advantage of the recent assembly of a draft genome and of the development of pan-genomic RAD-seq markers to decipher the demographic history of the differentiating populations and develop genome scans of adaptive differentiation. We showed that the SP diverged relatively recently, that is, few hundred years ago, and went through two successive bottlenecks followed by population size expansions, while the sympatric WP is currently experiencing a population decline. We identified outlier SNPs that were mapped onto the genome, but none were associated with the phenological shift or with subsequent adaptations. The strong genetic drift that occurred along the SP lineage certainly challenged our capacity to reveal functionally important loci.

KEYWORDS
demographic inference, detection of selection, phenology, pine processionary moth, population genomics, RAD-sequencing

1 | INTRODUCTION

Ecological speciation in sympatry, the process by which adaptation to contrasting ecological conditions drives the divergence of co-
occurring populations, has received growing attention in the last 12 years (Rundle & Nosil, 2005). The fate of diverging populations maintaining a certain level of gene flow, and the conditions in which speciation can still occur are central questions in evolutionary biology (Smadja & Butlin, 2011). A mechanism possibly causing sympatric speciation is allochronic differentiation, which occurs when differences in breeding time within a species lead to temporal associative mating and limit gene flow between early and late reproducers (Alexander & Bigelow, 1960). Isolation by time can further lead to adaptation by time (Hendry & Day, 2005) when divergent selection operates between contrasting environmental conditions encountered at the different breeding times. This process remains largely unexplored in the literature, but has been suggested to occur in a large array of organisms such as plants (Devaux & Lande, 2008; Savolainen et al., 2006; Weis et al., 2005), birds (Friesen et al., 2007), fishes (Limborg, Waples, Seeb, & Seeb, 2014), corals (Rosser, 2015, 2016) or insects (Santos, Burban, et al., 2011; Sota et al., 2013; Yamamoto, Beljaev, & Sota, 2016; Yamamoto & Sota, 2009, 2012). Many more examples probably remain to be discovered, and only nine study cases were identified in a recent review of 200 papers as examples of "true allochronic speciation" (Taylor & Friesen, 2017). To go beyond the description of such case studies and disentangle the evolutionary scenarios underlying allochronic differentiation, much remains to be done; in particular, the initial reduction in migration rate between the diverging populations and the underlying genomic mechanisms remains to be explored in most cases.

The recent advent of high-throughput genomic techniques as well as statistical advances for analysing large-scale data sets has opened unprecedented opportunities to address ecological, evolutionary and genetic questions in nonmodel organisms (Hassellmann, Ferretti, & Zayed, 2015). Genomewide data have been proven to be powerful for estimating the age of demographic events (McCoy, Garud, Kelley, Boggis, & Petrov, 2014), retrieving fine-scale population genetic structures (Sulkin, Gagnaire, Bierne, & Charmantier, 2016) and identifying phylogeographic patterns (Derkarabetian, Burns, Starrett, & Hedin, 2016). More, even if studying wild populations of nonmodel organisms is still a major challenge, population genomic approaches have allowed identification of genomic regions underlying phenotypic characteristics or traits involved in local adaptation (Berdan, Mazzoni, Waurick, Roehr, & Mayer, 2015; Guo, DeFaveri, Sotelo, Nair, & Merliä, 2015; Hohenlohe, 2014). Here, we used population genomics in a nonmodel insect species to disentangle the evolutionary scenario of allochronic differentiation, followed by adaptation to new environmental conditions.

We focused on one of the few cases identified by Taylor and Friesen (2017) as a well-documented example of "true incipient allochronic speciation," namely the pine processionary moth Thaumetopoea pityocampa (Dennis & Schiffermüller). This species is a well-studied pest of pine trees over the Mediterranean basin. Its caterpillars bear urticating hair, causing public and animal health concern (Battisti, Holm, Fagrell, & Larsson, 2011; Battisti, Larsson, & Roques, 2017; Rodríguez-Mahillo et al., 2012). Briefly, T. pityocampa reproduces in summer and larval development occurs through autumn and winter all over its range. Reproduction immediately follows adult emergence, as adults have a very limited lifespan of 1–2 days. In 1997, a population of T. pityocampa showing a shift in phenology (reproduction in spring and larval development in summer) was discovered in the Mata Nacional de Leiria (MNL) in Portugal, where it co-occurred with individuals following the typical biological cycle (Pimentel et al., 2006; Santos et al., 2007). This unique shifted population is known as the "summer population" (SP) as opposed to all other known populations that are referred to as "winter populations" (WPs), in relation to the development time of the conspicuous larvae. The SP was initially restricted to a small area of the Mata Nacional and has been slowly expanding along the coast since then (Godefroid et al., 2016). Strikingly, all larval stages of the SP develop under radically different environmental conditions compared to the typical WPs, experiencing much higher temperatures that were so far supposed to be lethal to early larval stages (Huchon & Démolin, 1970; Santos, Paiva, Tavares, Kerdelhué, & Branco, 2011). Understanding the scenario of this divergence is thus of interest in the context of current climate warming.

Previous studies have brought significant preliminary knowledge about the genetic and ecological characteristics of the peculiar SP. Analysis of a fragment of the mitochondrial COI gene and of the ITS1 region showed a high sequence similarity between the SP and the sympatric WP, which suggested a local origin of the SP, while microsatellites revealed a high differentiation between the SP and all studied Iberian populations (Santos, Burban, et al., 2011; Santos et al., 2007). Moreover, a recent study showed that some individuals belonging to the SP genetic cluster emerge during the WP reproductive season and are referred to as "LateSP individuals" (Burban et al., 2016). This study also documented signs of rare hybridization between the two allochronic populations. Consistently, hybrids between SP and WP individuals could be obtained in laboratory conditions, and the time of adult emergence (a proxy for breeding time) was shown to be highly heritable (Branco, Paiva, Santos, Burban, & Kerdelhué, 2017). These patterns suggested that the SP originated from the WP, following a phenological shift of a few individuals, and that gene flow between the SP and the WP is now highly reduced but not absent. Yet, the population genetic data relied on a limited number of microsatellite markers and did not allow us to characterize the successive stages of the divergence between the SP and the WP.

The objectives of the present work were to uncover major characteristics of this prime example of allochronic differentiation and significantly move towards the fulfilment of the criteria proposed by Taylor and Friesen (2017) by deciphering the evolutionary history of the SP and characterizing its different stages. In particular, we aimed at (i) inferring the timing of the divergence, (ii) measuring the migration rate between diverging populations at different stages to determine whether the differentiation occurred in the presence or absence of gene flow; (iii) determining the extent of population size changes, in particular to decipher if the SP experienced a strong bottleneck during the primary divergence step; and (iv) characterizing genomic regions possibly involved in the phenological shift and
subsequent adaptations. To achieve these aims, we took advantage of RAD-seq technology (Baird et al., 2008; Davey & Blaxter, 2011) and the recent release of a first draft genome for *T. pityocampa* (Gschlössl et al., Unpublished data) to obtain a large number of informative loci genotyped in the SP and in two WPs occurring in the same region. We used these loci to explore complex demographic scenarios including drift, migration and variation in population sizes, and to perform genomewide scans for signatures of selection. We could thereby successfully disentangle the main characteristics of the ongoing allochronic differentiation process.

2 | MATERIALS AND METHODS

2.1 | Biological material

A total of 180 individuals (adults or larvae) of *T. pityocampa* were collected in Portugal between May 2008 and September 2010 following Santos, Burban, et al., 2011. These individuals originated from three distinct populations or sampling sites: two were collected in the MNL (39°47′N 8°58′W) and corresponded to the winter and summer populations from Leiria (referred to as LWP and LSP), and one winter population was collected in the Setubal peninsula, near Apostiça (38°34′N 9°07′W), ca. 150 km south from Leiria, at the same elevation and longitude, and was hereafter denoted as AWP. All individuals were sampled from the host plant *Pinus pinaster* Aiton.

Forty L5 larvae (i.e., fifth larval stage) belonging to the AWP were collected in December 2010; 40 males, 10 females and 20 L5 larvae belonging to the LWP were sampled in 2008–2010, and 60 males and 10 females belonging the LSP were sampled in 2008–2010. For the LSP and LWP, we used two subsamples in each case. The first one included individuals assigned to the winter or the summer population based on the phenology observed in the field following Santos, Burban et al. (2011) (*n* = 40 for each population, subsamples referred to as LSP1 and LWP1). The second subsample gathered males caught with pheromone traps and previously genotyped using 17 microsatellite loci, from which we excluded the individuals assigned as LateSP, F1 and F2 following Burban et al. (2016); these subsamples (*n* = 30 in each population) will be referred to as LSP2 and LWP2. The exact sampling design is described in Table 1.

Table 1. The DNAs of each of the three sampled populations (all the 40 LSP1, all the 40 LWP1 and all the 40 AWP individuals, respectively) were pooled, and each population pool was barcoded with three different barcodes (nine barcodes in total, library #4). The RAD libraries #1 to #4 were then combined and PE sequenced (2 × 101 bp) on two Illumina HiSeq2000 lanes in the Edinburgh Genomics facility.

To replicate the experiment using only individuals genetically assigned to the LSP and LWP clusters following Burban et al. (2016), we further constructed 1 LSP2 and 1 LWP2 RAD libraries. Library #5 was constructed using 20 LWP2 males individually barcoded and a pool of all the 30 LWP2 DNAs that was identified with ten different barcodes. Finally, library #6 was the counterpart of library #5 for the LSP2 batch. Libraries #5 and #6 were each PE sequenced (2 × 101 pb) on a single Illumina HiSeq2000 lane on The Edinburgh Genomics facility.

2.2.2 | Bioinformatic analyses

Reads were first demultiplexed according to their barcode into individual and pool sequences using the default options of the `process_radtags` program of the `STACKS` package (version 0.99994) (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), including -q to remove poor-quality reads. PCR duplicates were further discarded using the `clone_filter` program (`STACKS` version 0.99994). The remaining reads were trimmed by removing the first five bases and keeping the next 90 bases for reads 1 and keeping the first 95 bases for reads 2. The reads originating from the same population identified with different barcodes or from the same individual run on different lanes were merged to increase coverage. We decided to discard three LSP1 and two LWP1 individuals from further analyses because their final coverage was too low, hence resulting in 75 genotyped individuals. The
number of remaining PE reads for the Ind-Seq data sets varied from 996,796 to 20,480,652 with a total of ca. 304 millions (111,015,950; 81,905,824; 66,817,320 and 44,207,850 PE reads for the LSP1; LWP1; LWP2 and LSP2 individuals, respectively). Similarly, ca. 300 millions of PE reads were kept for the Pool-Seq data sets (59,692,064; 86,867,660; 91,755,446; 37,102,870 and 23,485,726, PE reads for the AWP; LSP1; LWP1; LSP2 and LWP2 pools, respectively). RAD-seq PE reads were then mapped to the indexed Tpit-SP V1 assembly (Gschloessl et al., Unpublished data) using the bwa aln and bwa sampe commands of the BWA 0.6.2 program (Li & Durbin, 2009) with default options to generate bam files for each of the 75 remaining individuals (38 LWP, i.e., 18 LWP1 + 20 LWP2; and 37 LSP, i.e., 17 LSP1 + 20 LSP2) and the five pool samples (Table 1). Between 56.1% and 67.3% RAD sequences were mapped and properly paired onto the genome for the different data sets (mean insert size: 286 bp).

2.2.3 | Generation of the Ind-Seq SNP data set (gIS)

The RAD Ind-Seq bam files were processed using the mpileup command of SAMTOOLS 0.1.19 (Li et al., 2009) and the same default options as above to obtain LWP and LSP mpileup files. SNP and genotype calling were then performed separately for each of these two files using the bcftools view command, and the resulting vcf files were merged using the vcf-merge program from the VCFTOOLS 0.1.12 package (Danecek et al., 2011) after filtering variants using the vcfutils.pl varFilter command from the SAMTOOLS suite with default options and \(-w 5\) \(-d 200\). Because of the high heterogeneity in the observed within-SNP and within-individual read coverages, we performed additional filtering steps to obtain a genotyping data set as comprehensive as possible. First, all the genotypes with a read coverage \(DP < 5\) or \(DP > 1,000\) or a Phred quality \(GQ < 20\) were treated as missing data. The resulting number of genotype calls varied between 7,272 and 180,600 (with a median of 38,130). We thus decided to focus on the 40 individuals (28 LSP, i.e., 16 LSP1 + 12 LSP2; and 12 LWP, i.e., 10 LWP1 + 2 LWP2) with more than 35,000 genotype calls. We then discarded all the SNPs that were called on less than 90% of these 40 individuals leading to a total of 6,488 remaining SNPs. The resulting Ind-Seq genotyping data set—hereafter referred to as gIS—had the following characteristics: (i) the individual genotyping call rate varied between 83.6% and 99.9% with a median equal to 96.3%; (ii) the individual mean read coverage varied between 10.8 and 72.1 with a median equal to 16.1; (iii) the SNP genotyping call rate varied between 92.5% and 100% with a median equal to 95.0%; and (iv) the SNP minor allele frequency varied between 0.012% and 0.5% with a median equal to 0.14%.

2.2.4 | Generation of the Pool-Seq SNP data sets (rPS and pPS)

The five RAD Pool-Seq bam files were processed using the mpileup command of SAMTOOLS with default options and \(-d 5000\) \(-a 20\). The resulting file was further processed using a custom awk script to compute read counts for each alternative base after discarding bases with a BAQ quality score <25. A position was then considered as variable if (i) it had a coverage of more than 20 and less than \(c_i^{(\text{max})}\) reads in each population \(i\), where \(c_i^{(\text{max})}\) represented the 95th percentile of the coverage of all positions for population \(i\); (ii) only two different bases were observed across all the five pools; and (iii) the minor allele was represented by at least one read in two different pool samples.

The final data read count for the Pool-Seq data set—hereafter referred to as the rPS data set—consisted of 58,210 SNPs with mean (median; max) coverage equal to 34.96 (32; 67) for the LWP2 pool; 53.71 (50; 105) for the LSP2 pool; 76.21 (72; 164) for the AWP pool; 122.9 (116; 244) for the LWP1 pool; and 124.6 (121; 253) for the LSP1 pool, respectively.

For applications requiring allele count data (joint PCA of individual and Pool-Seq data, computation of the SFS, see below), we used the following approach. Let \(a_j\) represent the number of reads of the reference allele and \(c_j\) the coverage for SNP \(i\) in population (pool) \(j\) with haploid sample size \(n_j\). We further denote similarly \(y_j\) the allele count for the reference allele in the sample and \(n_j\) the haploid sample size \(n_j \leq n_j\) for SNP \(i\) in population (pool) \(j\). The pPS data set consists of the \(y_j\)’s and \(n_j\)’s, which were computed as follows:

1. if \(c_j \leq n_j\) then \(y_j = c_j\) and \(y_j = a_j\)
2. if \(c_j > n_j\) then \(y_j = n_j\) and \(y_j = a_j\)

\(a_j = 0 \) if \(n_j \leq a_j = c_j\) then \(y_j = a_j\);
\(a_j > c_j \) then \(y_j = (n_j - 1) \land (1 \lor (\forall i \in \Omega \ni a_j / c_j))\)

where \(\land\) and \(\lor\) stand for the maximum and the minimum, respectively. Note that formally (2) provides the maximum-likelihood estimate of the \(y_j\)’s under the assumption that the \(a_j\)’s follow a binomial distribution \(a_j = \text{Bin}(y_j/n_j, n_j)\). This approximation thus amounts in assuming equal contribution of each individual of the pool to the Pool-Seq read data.

2.3 | Population genetic diversity and structure

2.3.1 | Estimation of \(F_{ST}\) from Pool-Seq data

Pairwise and across populations \(F_{ST}\) were estimated using the estimator by Weir and Cockerham (1984) from the pPS data set. Even though this standard estimator was developed to measure differentiation from allele count data and therefore should be used cautiously when considering Pool-Seq data, the inherent biases are expected to be limited here given the haploid pool size, sequencing coverage and level of differentiation of the populations under study (Hivert V, Gautier M & Vitalis R, personal communication, September 2017).

2.3.2 | Estimation and visualization of the scaled covariance matrix of the population allele frequencies

To further assess the overall structuring of genetic diversity, we estimated the scaled covariance matrix of allele frequencies (\(\Theta\) across
the five samples using the software BayPass (Gautier, 2015) with default options. When applied to read count data (rPS data), the Bayesian model underlying BayPass provides an accurate estimate of \( \Omega \) by integrating over the unobserved allele count estimation. An eigenvalue decomposition of the resulting \( \Omega \) matrix was further performed using the \( \text{svd}() \) function to represent its major axis of variation. This latter approach amounts to performing a PCA that accounts rigorously for the specificities of the Pool-Seq data in the estimation of the covariance matrix.

### 2.3.3 Joint principal component analyses of individual (gIS) and pool-Seq (pPS) data

A total of 742 SNPs were in common between the individual gIS and the pPS pool data sets. We combined both data sets to obtain a matrix \( X = \{x_{ij}\} \) (742 SNPs \( \times 45 \) columns) of allele counts in 40 diploid individual samples \( \{n_j = 2\} \) and five pool samples \( \{n_{41} = n_{\text{AWP}} = 80\} \) and \( n_j = n_{\text{LSP1}} = n_{\text{LWP1}} = n_{\text{LSP2}} = n_{\text{LWP2}} = 60 \) for \( j = 42-45 \) resulting in a total of 360 haploid individuals. To account for the differences in sample size, we defined a SNP weight vector \( w = \{w_j\} \) where \( w_j = 1/180 \) for \( j = 1-40 \), \( w_{41} = n_{\text{AWP}}/180 = 40/180 \) and \( w_j = w_{\text{LSP1}} = w_{\text{LWP1}} = w_{\text{LSP2}} = w_{\text{LWP2}} = 30/180 \) for \( j = 42-45 \). Then we computed the (observed) allele frequencies \( f_i = x_{ij}/n_j \) for each SNP \( i \) and sample \( j \), and the overall mean weighted allele frequency \( \pi_i = \Sigma w X_{ij} \) for each SNP \( i \). Note that \( f_i \) was set to \( p_i \) when \( x_{ij} \) was missing. Finally, we computed the standardized allele frequency matrix \( M = \{m_{ij}\} \) where \( m_{ij} = (x_{ij} - p_i)/(p_i(1-p_i)) \).

A weighted principal component analysis (PCA) was then carried out based on the matrix \( M \) and using \( w \) as a (row) weight vector with the \texttt{dudi.pca()} function of the R package ade4 (Chessel, Dufour, & Thioulouse, 2004).

### 2.4 Demographic inferences

#### 2.4.1 Three-population tests of admixture

\( F_3 \) statistics provide a formal test for population admixture in three-population trees (Patterson et al., 2012). A significantly negative \( F_3 \) statistics for a \((P1; P2, P3)\) configuration supports an admixed origin of population \( P1 \) with two ancestral source populations related to \( P2 \) and \( P3 \), respectively. Note however that the reverse is not necessarily true, for example, the \( F_3 \) statistics might not be significantly negative in this same configuration if \( P1 \) experienced strong drift after the admixture event. \( F_3 \)-based tests were carried out for the possible topologies using the rPS POOL-SEQ data set. To account for the additional sampling level introduced in POOL-SEQ experiments (i.e., the sampling of read sequences in the DNA pool), the following unbiased estimator relying on read count data and haploid pool sizes was used:

\[
\hat{F}_3(A, (B, C)) = \frac{1}{T} \left[ \frac{1}{d} \sum_{d \in D} \left( \hat{\alpha}_d(A) + \hat{\beta}_d(B, C) - \hat{\beta}_d(A, B) - \hat{\beta}_d(A, C) \right) \right]
\]

with:

1. \[
\hat{\alpha}_d(P) = \frac{1}{n_P - 1} \left( a_{dP} (\alpha_{dP} - 1) - a_{dP} \right) \frac{\alpha_{dP}}{c_{dP} (c_{dP} - 1)}
\]

2. \[
\hat{\beta}_d((P, Q)) = \frac{a_{dP} a_{dQ}}{c_{dP} c_{dQ}}
\]

where for SNP \( d \) and \( a_{dP} \) (resp. \( a_{dQ} \)) represents the number of reads of the reference allele and \( c_{dP} \) (resp. \( c_{dQ} \)) the coverage in population \( P \) (resp. \( Q \)) with haploid sample size \( n_P \). To assess the significance of the departure of each statistic to the null hypothesis \((F = 0)\), Z-scores were computed as the ratio of the \( F_3 \) mean to its standard deviation both estimated over 5,000 bootstrap samples.

#### 2.4.2 Estimation of tree topology and divergence times using KIMTREE

For a given tree topology, we estimated divergence times using KIMTREE 1.3 (Gautier & Vitalis, 2013), with the standard MCMC parameters recommended in the user manual. KIMTREE is a hierarchical Bayesian model where the allele frequencies are modelled along each branch of a population tree using Kimura’s time-dependent diffusion approximation for genetic drift (Kimura, 1964). The support of the different topologies was assessed using a deviance information criterion (DIC) computed as described in Gautier and Vitalis (2013), and up to a constant term, with slight modifications for POOL-SEQ data:

\[
\text{DIC} = \frac{2}{T} \sum_{t=1}^T \sum_{i=1}^I \sum_{j=1}^J \sum_{l=1}^L -2 \log \left[ \left( c_j \right) \left( y_i(t) \right)^{q_i} \left( 1 - y_i(t) \right)^{q_i - q_j} \right] - \sum_{i=1}^I \sum_{j=1}^J -2 \log \left[ \left( c_j \right) \left( y_i(t) \right)^{q_i} \left( 1 - y_i(t) \right)^{q_i - q_j} \right]
\]

where \( y_i(t) \) represents the sampled allele count value for SNP \( i \) in pool \( j \) at the \( t \)th MCMC iteration (out of \( T \)) and \( \hat{y}_i \) the posterior mean of the corresponding allele count computed over the \( T \) MCMC samples.

#### 2.4.3 Inference of complex demographic histories using FASTSIMCOAL

To explore more complex demographic scenarios, we analysed the joint site frequency spectrum (SFS) of the three sampled populations using the approach of Nielsen (2000) implemented in FASTSIMCOAL 2.5.2.21 (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, & Foll, 2013). This approach uses coalescent simulations to infer the likelihood of the observed SFS under any demographic model and performs well even in situations where the events are recent (Excoffier et al., 2013). The analyses were run on the folded SFS, that is, using the observed counts of the minor allele, obtained from the rPS data sets of the LSP2, LWP2 and AWP samples (LSP2 and LWP2 were chosen based on the \( F_3 \) statistics, see Result section) with a pool size of 30, that is, the haploid size of the smallest pool. We directly
estimated the scaled parameters of the models in a coalescent or a diffusion timescale (i.e., 2Nμ, 2Nm, T/N and μT) and then inferred the canonical parameters (divergence time T expressed in generations, migration rates and population sizes expressed in number of genes N, i.e., twice the number of diploid individuals) using the mutation rate μ = 2.9 × 10⁻⁹ mutations per generation per SNP, as recently estimated for the Lepidoptera Heliconius melpomene (Keightley et al., 2015). Note that this estimate of mutation rate is lower than that of Drosophila melanogaster (Haag-Liautard et al., 2007) and is expected to be more appropriate for our Lepidoptera model species. Estimated divergence times thus tend to be older than if we had used the Drosophila mutation rate.

We first considered a simple model of pure divergence and drift (DivDrift), which corresponds to the KimTree model, for the three possible topologies. This analysis allowed us to compare inferences of tree topologies and scaled divergence times obtained with two different methods and therefore to check that the FastSimCoal algorithm performed well when used with the pPS data set. We could identify the most likely topology in this simple model and then further increase step-by-step the complexity of the scenario. First, we incorporated past fluctuations in population size (DivDriftVar); second, we allowed migration between all populations (DivDriftMig); and third, we considered both variations in population size and migration (DivDriftVarMig). All models were compared using the Akaike information criteria (AIC). Inferences of the canonical and/or scaled parameters were only considered for the simple DivDrift model for comparison with KimTree and for the best-supported model. Estimated divergence times thus tend to be older than if we had used the Drosophila mutation rate.

Using the posterior mean of these parameters, the GenoDivergence software was run to identify the most likely topology in this simple model and then further investigate possible topologies. This analysis allowed us to compare inferences of the canonical parameters (divergence time T, scaled migration rate Nm, and population size C0) and the scaled migration rate Nm = 2.9 × 10⁻⁹ mutations per generation per SNP, as recently estimated for the Lepidoptera Heliconius melpomene (Keightley et al., 2015). Note that this estimate of mutation rate is lower than that of Drosophila melanogaster (Haag-Liautard et al., 2007) and is expected to be more appropriate for our Lepidoptera model species.

We first considered a simple model of pure divergence and drift (DivDrift), which corresponds to the KimTree model, for the three possible topologies. This analysis allowed us to compare inferences of tree topologies and scaled divergence times obtained with two different methods and therefore to check that the FastSimCoal algorithm performed well when used with the pPS data set. We could identify the most likely topology in this simple model and then further increase step-by-step the complexity of the scenario. First, we incorporated past fluctuations in population size (DivDriftVar); second, we allowed migration between all populations (DivDriftMig); and third, we considered both variations in population size and migration (DivDriftVarMig). All models were compared using the Akaike information criteria (AIC). Inferences of the canonical and/or scaled parameters were only considered for the simple DivDrift model for comparison with KimTree and for the best-supported model. Estimated divergence times thus tend to be older than if we had used the Drosophila mutation rate.

Whole genome scans for adaptive differentiation were carried out by looking for overly differentiated SNPs using both Selestim version 1.1.3 (Vitalis, Gautier, Dawson, & Beaumont, 2014) and Baypass version 2.1 (Gautier, 2015) that are both handling Pool-Seq data. Selestim is based on a diffusion approximation for the distribution of allele frequencies in a subdivided population, which explicitly accounts for selection. In particular, Selestim assumes that each and every locus is targeted by selection to some extent, and estimates the strength of selection at each locus in each population. For each analysis, twenty-five short pilot runs (1,000 iterations each) were set to adjust the proposal distributions for each model parameter and, after a 100,000 burn-in period, 100,000 updating steps were performed with a thinning interval of 40 steps. Candidate markers under selection were selected on the basis of the distance between the posterior distribution for the locus-specific coefficient of selection and a “centring distribution” derived from the distribution of a genomewide parameter that accounts for the among-locus variation in selection strength.

Selestim uses the Kullback-Leibler divergence (KLD) as a distance between the two distributions, which is calibrated using simulations from a posterior predictive distribution based on the observed data (Vitalis et al., 2014). Hereafter, we report candidate markers with KLD values above the 99.9% quantile of the so-obtained empirical distribution of KLD.

In Baypass, we identified candidate markers using the XTX differentiation measure ( Günther & Coop, 2013 ). This metrics might be viewed as a SNP-specific FST that explicitly corrects for the scaled covariance of population allele frequencies (matrix Ω), making it robust to the unknown demographic history relating the populations. The XTX was estimated using default options of Baypass. Pairwise correlations of the XTX estimates across ten independent runs were all found to be above 0.995 demonstrating the stability of the estimates. As described in Gautier (2015), the XTX was calibrated based on a posterior predictive distribution obtained by analysing a pseudo-observed data set of 250,000 SNPs generated under the inference model with hyperparameters fixed to their respective posterior means as estimated from the analysis of the original data. Hereafter, we report candidate markers with XTX values above the 99.9% quantile of the so-obtained empirical distribution of XTX. To identify the population of origin of the signal for overly differentiated SNPs, we examined the posterior means of the standardized population allele frequencies defined for each SNP i and population j as:

\[
X_i = \{x_{ij}\}_{j=1}^J = \frac{1}{\sqrt{\pi_i(1 - \pi_i)}} \Gamma^{-1} x_i
\]

where \(x_i\) represents the (unobserved) vector of population allele frequencies, \(\pi_i\) represents the across population allele frequency, \(\pi_i\) represents the across population allele frequency, and \(\Gamma\) represents the Cholesky decomposition matrix of \(\Omega\) (i.e., \(\Omega = \Gamma \Gamma^T\)). Although the standardized allele frequencies \(x_i\) are expected to be independent and identically normally distributed under the null model ( Günther & Coop, 2013 ), the Bayesian hierarchical model-based estimation procedure leads to shrink their estimated posterior mean. As a result, they are each calibrated as the XTX, that is, using their respective empirical distribution obtained from the analysis of the pseudo-observed data set described above.

3 | RESULTS

3.1 | Population genetic diversity and structure

The multilocus \(F_{ST}\) across the five samples was equal to 0.259, while pairwise population \(F_{ST}\) varied from 0.038 for the (LSP1; LSP2) pair to 0.374 for the (AWP; LSP2) pair (Table 2). The LWP1 and LWP2 samples appeared differentiated, with a pairwise \(F_{ST}\) equal to 0.068. Nevertheless, both the LWP1 and LWP2 samples were found closer to the AWP ( \(F_{ST}\) equal to 0.095 and 0.125 for the (AWP; LWP1) and (AWP; LWP2) pairs, respectively) than to the LSP ( \(F_{ST}\) ranging from 0.302 to 0.368 depending on the subsamples representing LSP and LWP).
TABLE 2 pairwise $F_{ST}$ estimates between the analysed pools

|       | AWP | LSP1 | LSP2 | LWP1 |
|-------|-----|------|------|------|
| AWP   | 0.369 |      |      |      |
| LSP1  | 0.369 | 0.038|      |      |
| LSP2  | 0.095 | 0.302| 0.307|      |
| LWP1  | 0.125 | 0.368| 0.362| 0.068|
| LWP2  |      |      |      |      |

We estimated the scaled covariance matrix of allele frequencies $\Omega$ across the five samples using BAYPASS and performed an eigenvalue decomposition of that matrix, which results in a principal component analysis accounting for the specificities of the Pool-Seq data. As shown in Figure 1a, the first axis of variation (PC1) accounted for 93.5% of the total genetic variation and separated the samples according to the phylogeography of their underlying population (i.e., LSP1 and LSP2 vs. LWP1, LWP2 and AWP). The second axis of variation (PC2) that only accounted for 4.17% of the total variation was associated with a geographic gradient as it separated the Leiria samples (LSP1, LSP2, LWP1 and LWP2) from the Apostiça sample (AWP). Importantly, the coordinates of both the LSP1 and LWP1 samples on PC1 were found closer to the origin than their corresponding LSP2 and LWP2 counterparts. This result suggests the presence of LateSP individuals and/or hybrids in either LSP1 or LWP1 or both.

This latter result was supported by the joint analysis of a subset of the Pool-Seq data together with the 28 LSP and 12 LWP genotyped individuals that had more than 35,000 genotype calls. Figure 1b shows the first factorial plan of a joint PCA performed on 742 SNPs that were in common between the Pool-Seq data set (pPS) and the individual data set (gIS). Although the number of SNPs was lower and the data for pool samples were projected onto their corresponding haploid sample size, the overall picture displayed in Figure 1b was qualitatively similar to that of Figure 1a. Interestingly, based on their coordinates on PC1, at least three of the 12 genotyped LWP1 individuals appeared to be either LateSP or introgressed individuals. When ignoring these three individuals, the coordinates of LWP individuals on PC1 were very close to that of the LWP2 pool sample. Conversely, the PC1 coordinates for all LSP individuals remained close to those of both the LSP1 and LSP2 pool samples. All results (PCA and BAYPASS) thus suggested a higher variability across the LWP samples than across the LSP ones. They also revealed that some LateSP and introgressed individuals were included in the LWP1 pool that contained individuals that were only phenotypically assigned to their “pheno-logical” population.

3.2 | Demographic inference

3.2.1 | $F_3$-based tests of admixture

Three-population tests were carried out for all the 30 possible configurations among the five pool samples (Table S2). Six configurations resulted in significant negative $F_{ST}$ statistics. They corresponded to the four configurations that tested the LWP1 sample against another WP sample (AWP or LWP2) and a LSP sample (either LSP1 or LSP2) as source populations: (i) (LWP1; AWP, LSP1) with $Z = -9.02$; (ii) (LWP1; AWP, LSP2) with $Z = -9.45$; (iii) (LWP1; LSP1, LWP2) with $Z = -17.8$; and (iv) (LWP1; LSP2, LWP2) with $Z = -11.7$. This result confirmed the inclusion of LateSP individuals in the LWP1 pool, as suggested by the PCA. The two other configurations displaying significantly negative $F_{ST}$ statistics tested the LSP1 sample against the LSP2 sample and either the LWP1 or AWP as samples representative of the WP: (i) (LSP1; AWP, LSP2) with $Z = -4.00$; (ii) (LSP1; LSP2, LWP1) with $Z = -3.57$. On the contrary, considering the LWP2 sample as representative of the LWP did not result in a significantly negative $F_{ST}$ (Table S2). In that case, the signal of admixture might be hidden by the stronger drift in LSP1. The $F_{ST}$ of the (LWP2, LSP1) pair being higher than that of the (LWP1, LSP1) pair. We here recall that the results of the $F_{ST}$ test should be interpreted only when significant. As both the PCA and $F_{ST}$ statistics suggested that the LWP1 pool probably contained LateSP individuals, we further used only the LSP2 and LWP2 samples as representing the LSP and LWP to infer the demographic history of the LSP.

FIGURE 1 Principal component analysis of gene frequencies across the AWP, LSP1, LSP2, LWP1 and LWP2 pool samples. (a) Plot of the pool sample coordinates on the first two axes of variation of $\Omega$, the scaled covariance matrix of allele frequencies across the five pool samples. The matrix $\Omega$ was estimated with BAYPASS (Gautier, 2015) using read count data (rPS) available on 58210 SNPs. (b) First factorial plan of the joint PCA performed on the projected allele count data (pPS) for the five pool samples together with genotyping data for 28 LSP and 12 LWP individuals. The combined data set consisted of 742 SNPs [Colour figure can be viewed at wileyonlinelibrary.com]
3.2.2 | Inferring the tree topology and divergence times under various scenarios

We first ran KIMTREE on the Pool-Seq rPS data to compare the four possible topologies relating the AWP, LSP (using LSP2 sample) and LWP (LWP2 sample) under a pure-drift model of divergence (Figure 2). The DIC gave the strongest support to the (LSP, (AWP, LWP)) tree (Figure 2). Interestingly, the branch length relating the LSP to the root population (ancestral to the winter and summer populations) revealed a strong signature of drift ($\tau_{\text{LSP}} = 0.383$).

We then analysed the joint SFS of the three populations using the estimated allele count data pPS for different demographic models. Considering a simple model of divergence and drift (Div-Drift), the best-supported tree according to the AIC corresponded to the best-supported tree obtained with KIMTREE (Table S3). In the following steps, we thus only considered the topology (LSP, (LWP, AWP)) (Table 3). SFS analyses under this model lead to precise estimates of scaled parameters such as population size ratios ($N_P/N_{\text{LSP}} = 7.9 [5.4; 9.9]$ and $N_{\text{WP}}/N_{\text{LSP}} = 26 [17; 35]$), and of the four scaled divergence times for the different branches of tree, that can be directly compared to those inferred from KIMTREE and appear to be highly consistent (Figure 2). Indeed, we estimated $T_P/N_{\text{LSP}} = 0.35 [0.34; 0.37]$ (to compare with $\tau_{\text{LSP}} = 0.383$ in Figure 2); $T_{\text{WP}}/N_{\text{LWP}} = 0.085 [0.081; 0.092]$ (to compare with $\tau_{\text{LWP}} = 0.099$ in Figure 2); $T_{\text{WP}}/N_{\text{AWP}} = 0.089 [0.081; 0.092]$ (to compare with $\tau_{\text{AWP}} = 0.117$ in Figure 2); and $(T_P - T_{\text{WP}})/N_{\text{WP}} = 0.11 [0.094; 0.12]$ (to compare with $\tau_{\text{P4}} = 0.107$ in Figure 2). This overall good agreement between the KIMTREE and SFS analyses suggest that estimating the SFS from the inferred allele counts (pPS data set) provides robust results (KIMTREE analyses being based on the read count rPS data set).

We further investigated more complex models by including migration between the populations (DivDriftMig model), variation in population sizes (DivDriftVar model) or both (DivDriftVarMig model). Comparison of AICs for these four demographic models showed that the data strongly supported the DivDriftVarMig model detailed in Figure 3 (Table S4). Most of the 24 canonical parameters of this latter model, that is, all population sizes and divergence times as well as some migration rates, were inferred with good precision (Table 3). The few exceptions concerned some migration rates for which CIs were relatively broad.

Overall, the SFS analyses suggested that the ancestral SP and WP diverged relatively recently, ca. 560 generations ago (with a confidence interval CI ranging from 448 to 2280), both experiencing a concomitant bottleneck. Then, the ancestral SP experienced a first expansion, followed by a second bottleneck ca. 69 (CI = 35–216) generations ago, and a second strong expansion until present. From the ancestral WP, LWP and AWP diverged ca. 207 generations ago (CI = 95–526). Note that age estimates depend on the mutation rate

**FIGURE 2** Comparisons under a pure-drift divergence model of three bifurcating tree (a–c) and a star phylogeny (d), relating the AWP, LSP (represented by the LSP2 pool sample) and LWP (represented by the LWP2 pool sample) using KIMTREE (Gautier & Vitalis, 2013). The tree with the highest support (smallest DIC) is represented in red and corresponds to ((AWP, LWP), LSP). For that tree, the posterior mean of the divergence time (measured on a diffusion timescale) is provided for each branch [Colour figure can be viewed at wileyonlinelibrary.com]
used, which is lower in *Heliconius* butterflies (Keightley et al., 2015) than in *Drosophila* (Haag-Liautard et al., 2007). LWP recently experienced a relatively severe contraction while AWP showed an expansion event. Accordingly, negative growth rates (corresponding to expansions in the coalescence analyses) were inferred for all but the LWP. Finally, inferred migration rates were relatively large for the pairs (AWP, LWP) and (LSP, LWP), with especially large values for the migration from LWP to AWP and to a lesser extent from LSP to LWP. On the contrary, inferred migration rates were lowest for the pair (LSP, AWP) as well as for the migration from SP to WP, that is, the ancestral populations.

### 3.3 Genome-scan for adaptive differentiation

We performed genome scans for adaptive differentiation across the three-population samples (AWP, LWP2 and LSP2) using both the *SELESTIM* and *BAYPASS* software packages. Of the 54,040 analysed SNPs (4,170 SNPs from the original rPS data set were discarded as monomorphic in the three analysed population pools), 12 were found outlier with *SELESTIM* and 73 with *BAYPASS*; 11 were in common between the two analyses (Figure 4; Figure S1 and Table S5). However, we found no locus presumably involved in the phenological shift or subsequent ecological adaptation in the LSP. Indeed, among the outlier SNPs, none displayed extreme value for either the population-specific coefficient of selection estimated with *SELESTIM*, or the standardized allele frequencies estimated with *BAYPASS* in the LSP2 sample only. Instead, most outliers displayed outstanding differentiation in both the LWP and LSP (data not shown).

We then mapped the outlier SNPs onto the recently obtained draft genome (Gschloessl et al., Unpublished data) and used the associated gene prediction and transcriptomic

---

**TABLE 3** Parameter point estimates and associated confidence intervals obtained for the best-supported demographic history (DivDriftVarMig with the population tree (LSP, (AWP, LWP))) from the SFS analysis. See Figure 3 for details.

| Type of parameter | Demographic parameter | "DivDrift" model | "DivDriftVarMig" model |
|-------------------|-----------------------|------------------|-------------------------|
| Effective population size | $N_{AWP}$ (0) | 2330 [2240; 1640 $10^3$] | 4.40 $10^3$ [2.85 $10^3$; 5.16 $10^3$] |
| | $N_{LSP}$ (0) | 2.16 $10^3$ [2.07 $10^3$; 4.05 $10^3$] | 6.63 $10^3$ [2.48 $10^3$; 1.63 $10^3$] |
| | $N_{LWP}$ (0) | 3790 [2930; 155 $10^3$] | 293 [129; 828] |
| | $N_{SP}$ ($T_{SP}$) | n.a. | 3.97 $10^3$ [1.23 $10^3$; 8.62 $10^3$] |
| | $N_{WP}$ ($T_{WP}$) | 7.16 $10^3$ [6.64 $10^3$; 1.38 $10^3$] | 4.97 $10^3$ [1.33 $10^3$; 1.28 $10^3$] |
| | $N_{P}$ ($T_{P}$) | 5.60 $10^3$ [5.43 $10^3$; 8.53 $10^3$] | 2.44 $10^3$ [2.17 $10^3$; 2.53 $10^3$] |
| | $N_{P}$ ($T_{ANC}$) | n.a. | 1.85 $10^3$ [9.83 $10^3$; 2.17 $10^3$] |
| | $N_{AWP}$ ($T_{WP}$) | n.a. | 43 [17; 121] |
| | $N_{LSP}$ ($T_{SP}$) | n.a. | 147 [52; 741] |
| | $N_{LWP}$ ($T_{WP}$) | n.a. | 1.42 $10^3$ [2.65 $10^3$; 4.85 $10^3$] |
| | $N_{SP}$ ($T_{P}$) | n.a. | 43 [35; 257] |
| | $N_{WP}$ ($T_{P}$) | n.a. | 129 [103; 974] |
| Divergence time (in generation) | $T_{SP}$ | n.a. | 69 [35; 216] |
| | $T_{WP}$ | 216 [190; 1380] | 207 [95; 526] |
| | $T_{P}$ | 7.59 $10^3$ [6.90 $10^3$; 1.38 $10^3$] | 560 [448; 2.28 $10^3$] |
| | $T_{ANC}$ | n.a. | 9.05 $10^3$ [8.79 $10^3$; 1.10 $10^3$] |
| Migration rate | $m_{(LSP \rightarrow AWP)}$ | n.a. | 5.36 $10^7$ [2.95 $10^5$; 9.38 $10^6$] |
| | $m_{(LWP \rightarrow AWP)}$ | n.a. | 5.15 $10^3$ [1.62 $10^3$; 1.04 $10^2$] |
| | $m_{(AWP \rightarrow LSP)}$ | n.a. | 2.40 $10^8$ [2.67 $10^6$; 9.86 $10^7$] |
| | $m_{(LWP \rightarrow LSP)}$ | n.a. | 9.69 $10^4$ [3.10 $10^4$; 1.91 $10^3$] |
| | $m_{(AWP \rightarrow LWP)}$ | n.a. | 4.29 $10^4$ [4.91 $10^4$; 1.10 $10^3$] |
| | $m_{(LSP \rightarrow LWP)}$ | n.a. | 1.11 $10^3$ [3.46 $10^2$; 2.53 $10^3$] |
| | $m_{(WP \rightarrow SP)}$ | n.a. | 2.29 $10^5$ [2.34 $10^5$; 2.55 $10^5$] |
| | $m_{(SP \rightarrow WP)}$ | n.a. | 8.72 $10^8$ [2.99 $10^5$; 8.50 $10^8$] |
| Growth rate | $G_{AWP}$ | n.a. | $-0.480 [-1.10; -0.174]$ |
| | $G_{LSP}$ | n.a. | $-0.764 [-1.89; -0.175]$ |
| | $G_{LWP}$ | n.a. | 0.258 [8.60 $10^2$; 0.551] |
| | $G_{SP}$ | n.a. | $-0.160 [-0.201; -2.79 $10^{-2}$] |
| | $G_{WP}$ | n.a. | $-0.201 [-0.247; -0.029]$ |
| | $G_{P}$ | n.a. | $-2.46 $10^5$ [-2.56 $10^5$; -2.22 $10^5$] |
resources to annotate the SNPs which fell within or near (<2,000 pb) a potential gene. The 74 SNPs mapped to 63 different scaffolds; seven of these SNPs were located within a gene (five in introns of four different predicted or reconstructed genes, two in exons of two genes), seven were located in the vicinity of five different predicted or reconstructed genes. Only two of the corresponding genes could be annotated, and corresponded to a transcription domain-associated protein of *Operophthera brumata* and an E3 ubiquitin-protein ligase RFWD2-like of the Pyralidae *Amyelois transitella*. These results are detailed in Tables S5 and S6.

4 | DISCUSSION

In this study, we analysed an illustrative example of “true allochronic differentiation” (sensu Taylor & Friesen, 2017) between sympatric populations of the pine processionary moth. Our results allowed us to decipher when and how the primary divergence occurred (bottleneck intensity, levels of gene flow), which allows us to propose hypotheses about the circumstances of the differentiation and the subsequent history of the populations.

4.1 | The primary divergence: a fairly recent allochronic event associated with a strong bottleneck and an abrupt disruption of gene flow

Tree-based analyses suggested that the phenologically shifted SP first diverged from the common ancestor of the two studied WPs, which differentiated more recently from one another. The long branch leading to the SP suggested that this population experienced very strong drift. The model was significantly improved by including changes in population sizes and migration between populations, suggesting that the demographic history associated with the allochronic event is relatively complex. We could infer in detail this evolutionary scenario. The common ancestor of the SP and the WP is supposed to have been present for a long time (estimated to 900,000 years), with large population sizes ($10^5$–$10^6$ reproducing individuals), which is consistent with the continuous occurrence of the pine processionary moth in the refugial areas of the Iberian Peninsula during the Ice Ages (Rousselet et al., 2010). The divergence of the SP was estimated to have occurred ca. 560 years ago, and it was associated with a very strong founder event (ancestral population size estimated to a few tens of individuals), while a bottleneck event occurred in the ancestral WP. One of the main questions about
sympatric differentiation is to know whether it occurred in the presence or absence of gene flow in the first steps of the divergence, and how migration evolved over time (Powell, Forbes, Hood, & Feder, 2014; Smadja & Butlin, 2011). The question of the levels of gene flow can shed light on the differentiation process and impact the possible fate of the diverging populations. In the particular case of allochronic differentiation, the shift in breeding time can occur progressively, an overlap in reproductive times of the incipient populations then maintaining gene flow (with some similarities between isolation by distance and isolation by time in this case, see Hendry & Day, 2005). Conversely, it can also appear as an abrupt phenological change that would immediately disrupt gene flow and lead to an “automatic” complete assortative mating, acting as an “automatic magic trait” sensu Servedio, Doorn, Kopp, Frame, and Nosil (2011). Our results showed that the first step of the differentiation occurred in a context of highly limited gene flow between the ancestral SP and WP (migration rate 10^{-5} to 10^{-8}). This corroborates the hypothesis of a sudden event of divergence, resulting in an immediate barrier to gene flow between the two incipient populations.

Allochrony can in some situations evolve as a by-product of another primary driver of speciation, such as host plant shift followed by alteration of breeding time to match with the new host's phenology (Powell et al., 2014). It is not possible from our results to rule out the hypothesis that the two populations primarily diverged due to other factors, and that allochrony evolved more recently and would now be the main differentiated trait. On the other hand, no host or habitat change is associated with the differentiation of the SP. The land was once covered by mixed forests and shrubs (AFN—Autoridade Florestal Nacional, 2012) and then sowed with _P. pinaster_ during the 13th and early 14th century. The divergence of the SP probably occurred after this large afforestation programme, which took place ca. 700 years ago, when _P. pinaster_ was already predominant in this region. We thus conclude that in the particular case of the pine processory moth, allochrony can still be hypothesized to be the initial driver of divergence. It is very likely that the periods of adult activity of the two diverging populations did not overlap in the early phases of their differentiation, immediately disrupting gene flow. A scenario of an initial mutation in key genes involved in seasonal rhythms or affecting diapause termination which first occurred by chance and drove the differentiation event in a very limited number of founder individuals can thus be favoured (Schluter, 2009), and would be consistent with the high heritability found in experimental rearing (Branco et al., 2017). This information is crucial for our understanding of the allochronic differentiation process.

We obtained a relatively recent estimate of the divergence time, but our results suggest that the SP was already present few hundred years before its discovery in 1997 (Santos, Burban, et al., 2011; Santos et al., 2007). No mention was found in the historical archives of the Mata Nacional de Leiria (MB, pers. obs.), even though these archives contain much information because the national park has been a major wood production area for more than seven centuries. Yet, it is also possible that the ancestral SP evolved in the same region, but outside the limits of the park, and remained undocumented in historical times. In a recent study, Godefroid et al. (2016) showed that the current distribution of the LSP is limited by the high summer temperature occurring elsewhere in Portugal, even though larvae of this population were proved to cope better with higher temperatures than larvae of Portuguese and French WPs (Santos, Paiva, et al., 2011). In the first steps of the differentiation, milder environmental conditions could have favoured the success of the diverging population. Interestingly, a period of colder climate known as the Little Ice Age occurred between years 1300 and 1900, including in Portugal, bringing favourable climatic conditions (Abrantes et al., 2005; Bartels-Jónsdóttir, Knudsen, Abrantes, Lebriero, & Eiriksson, 2006). Other phenotypic trait divergences between the SP and the WP were documented, with obvious adaptations to the environmental changes experienced by the SP eggs and larvae due to the shift in breeding time (Rocha et al., 2017; Santos, Paiva, Rocha, Kerdelhujet, & Branco, 2013; Santos, Paiva, et al., 2011), consistent with the concept of “adaptation by time” proposed by Hendry and Day (2005). Whether such phenotypic changes occurred over ca. 500 years or whether they occurred over some tens of generations as previously suggested (Santos, Burban, et al., 2011; Santos et al., 2007), these adaptations can still be considered as rapid.

### 4.2 | Recent demographic changes in the diverging population and increased recent gene flow

The best demographic model we obtained further suggested that the SP experienced a recent bottleneck ca. 70 years ago, which reduced the population to a few hundred reproducers at most. The SP then expanded again until its high current population size (between 25,000 and 100,000 individuals). The cause of this recent and drastic reduction in size is difficult to characterize and could be due to a local climatic or epidemiological event or to human activities (e.g., local habitat destruction, forest fire, management options). This bottleneck actually coincides with the recent establishment of intensive planning and forest management in the MNL. The first forest plan dates back from 1892 and was intensified during the 1960s, including management by clear-cuts and development of 120 km of forest roads (AFN—Autoridade Florestal Nacional, 2012). This major demographic event is consistent with the fact that the SP remained undetected in the recent history and was discovered only recently during a very severe and thus conspicuous outbreak in 1997 (Pimentel et al., 2006; Santos et al., 2007). Parallel to the SP history, our model also suggested a complex scenario for the studied WPs. AWP and LWP diverged ca. 200 years ago, with a very strong founder event in Apostiça as the estimated population size reached 43 individuals only. This event could be linked to human activities and to the deforestation process that occurred to provision wood and agricultural goods, which dramatically decreased forest land in the region of Lisbon (Devy-Vareta, 1985). This probably tended to fragment the PPM habitat and strongly reduced its populations in the vicinity of Lisbon. It is worth noting that in the recent years, the population size has strongly increased in Apostiça, which is consistent with the
recent afforestation activity, whereas the Leiria WP tended to decrease. Whether the decline of the WP observed in the MNL could be linked to possible competition between the sympatric summer and winter populations should now be tested. Monitoring tools could moreover allow us to determine if this is a long-term trend or if the local LWP would increase again. On the other hand, our results consistently show that AWP was closely related to LWP, and could not be used as an out-group as we initially planned. A thorough phylogeographic study of Portuguese and/or Iberian populations would now be helpful to understand the genetic structure of populations in this PPM clade (Rousselet et al., 2010) and to develop further demographic analyses.

To complete the picture, our results suggested that some gene flow currently occurs between existing populations. Not surprisingly, in the best demographic model, migration rates were maximal between the two WPs but they were also relatively high in both directions between the two sympatric LSP and LWP (ca. 10⁻³). This is consistent with the recent identification of few hybrid individuals by Burban et al. (2016). Interestingly, our results suggest that the level of gene flow between the sympatric populations is higher today than in earlier stages of differentiation. This could be explained by the recent geographic and demographic expansion observed in the SP, which could have increased the probability of contact and thus introgression between the two populations. We could also hypothesize that plasticity in reproductive time plays a role by allowing some degree of overlap in reproductive time between the two populations, which can possibly vary over time as occurs in some plants (Devaux & Lande, 2008). Some individuals belonging to the SP but emerging during the LWP reproductive season were recently identified with molecular markers (Burban et al., 2016). Such “LateSP” individuals can only be identified through genotyping and could also allow some introgression between the two populations. The results further showed that assigning individuals from their phenotype alone can lead to erroneous mixing of some LateSP individuals in the LWP1 pool, and that robust results could only be obtained when pooling genetically well-characterized individuals. Preliminary observations suggest that some of these LateSP correspond to the last-emerging SP individuals, that is, to events at the tail-end of the distribution of SP emergence time in July, during the early WP season. Other LateSP actually emerge very late, after the WP season, and could correspond to a dysfunction in diapause termination (Burban et al., 2016). The origins and fate of these categories of LateSP remain to be studied.

4.3 Identifying and interpreting signatures of selection

All of the SNPs identified by BAYPASS and SELESTIM as presumably under selection displayed population-specific signatures associated with both the LSP and the LWP, which did not allow us to clearly identify a pattern linked to the phenotypic evolution of the SP. It is likely that the strong drift experienced by the SP and the high level of differentiation between the SP and both LWP and AWP ($F_{ST} > 0.3$) impedes optimal use of genomic scans of adaptation. A similar challenge in revealing functionally important loci due to a stronger than expected background differentiation was encountered by Lozier, Jackson, Dillon, and Strange (2016) in their study of Bombus colour patterns. Moreover, even if RAD-seq was proved to be a powerful approach to easily develop population genomic studies for nonmodel organisms, the technique only allows us to analyse a reduced proportion of the genome, which increases the likelihood of missing the genomic region truly targeted by selection (Lowry et al., 2017; but see McKinney, Larson, Seeb, & Seeb, 2017). Our study also pointed a major challenge in arthropod genomics, which is the low proportion of functionally annotated genes. We could annotate only two of the genes in the vicinity of the detected candidate SNP, which strongly limits the functional interpretation of the results. Moreover, the draft genome currently available for T. pityocampa has low scaffold sizes (Gschoessl et al., Unpublished data), which explains why most of the identified SNP were found in different genomic fragments. Improving the genome assembly will greatly increase our analysing capacities.

4.4 Perspectives and future directions

Several studies have recently identified candidate genes involved in circadian and seasonal rhythms and in diapause termination, and their roles and interactions are increasingly understood (Denlinger, 2002; Derks et al., 2015; Wadsworth & Dopman, 2015). In particular, there is increasing evidence that genes involved in circadian rhythms are also involved in reproductive cycles (Fuchikawa et al., 2010; Levy, Kozak, Wadsworth, Coates, & Dopman, 2015; Ragland, Egan, Feder, Berlocher, & Hahn, 2011; Ragland & Keep, 2017). One possible approach will be to target those genes both to resequence them in the SP and WP and possibly identify sequence polymorphisms and to determine whether they are differentially expressed in the allochronic populations at key stages of the development. An alternative approach could be QTL-mapping, which has proved to be a successful strategy in a number of studies (e.g., Alem et al., 2013; Franchini et al., 2014). It is however expected to be tedious in the particular example of the pine processionary moth for which rearing in experimental conditions is a difficult task due to a high mortality, the urticating nature of its larvae and the obligate 1-year generation time (Berardi, Branco, Paiva, Santos, & Battisti, 2015; Branco et al., 2017; Rocha et al., 2017).

ACKNOWLEDGEMENTS

We thank H. Santos for sampling the PPM individuals analysed in this article. We are grateful to B. Gschoessl who facilitated access to the genome assembly and provided the SNP annotations, and to A. Estoup for fruitful discussions in the early steps of the project. The authors also wish to thank Pierre-Alexandre Gagnaire for the RAD adaptors and Hélène Holota from TAGC facility for the Covaris sonication of the RAD libraries. This work was supported by a grant from the French National Research Agency (ANR-10-JCJC-1705-01).
REFERENCES

Abrantes, F., Lebreiro, S., Rodrigues, T., Gil, I., Bartels-Jonsdottir, H., Oliveira, P., ... Grimalt, J. O. (2005). Shallow-marine sediment cores record climate variability and earthquake activity off Lisbon (Portugal) for the last 2000 years. Quaternary Science Reviews, 24, 2477–2494. https://doi.org/10.1016/j.quascirev.2004.04.009

AFN · Autoridade Florestal Nacional (2012). Estratégia para a gestão das matas nacionais. Lisboa, Portugal: Relatório Direção Nacional de Gestão Florestal.

Alem, S., Streiff, R., Courtois, B., Zenboudjii, S., Limousin, D., & Greenfield, M. D. (2013). Genetic architecture of sensory exploitation: QTL mapping of female and male receiver traits in an acoustic moth. Journal of Evolutionary Biology, 26(12), 2581–2596. https://doi.org/10.1111/jeb.12252

Alexander, R. D., & Bigelow, R. S. (1960). Allochronic speciation in field crickets, and a new species, Acheta veletis. Evolution, 14(3), 334–346. https://doi.org/10.1111/j.1558-5646.1960.tb03095.x

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), e3376. https://doi.org/10.1371/journal.pone.0003376

Bartels-Jonsdottir, H. B., Knudsen, K. L., Abrantes, F., Lebreiro, S., & Eiriksson, J. (2006). Climate variability during the last 2000 years in the Tagus Prodelta, western Iberian Margin: Benthic foraminifera and stable isotopes. Marine Micropaleontology, 59, 83–103. https://doi.org/10.1016/j.marmicro.2006.01.002

Battistì, A., Holm, G., Faggrell, B., & Larsson, S. (2011). Urticating hairs in arthropods: Their nature and medical significance. Annual Review of Entomology, 56(1), 203–220. https://doi.org/10.1146/annurev-ento-120709-144844

Battistì, A., Larsson, S., & Roques, A. (2017). Processionary moths and associated urtication risk: Global change-driven effects. Annual Review of Entomology, 62(1), 323–342. https://doi.org/10.1146/annurev-ento-031616-034918

Berardi, L., Branco, M., Paiva, M.-R., Santos, H., & Battistì, A. (2015). Development time plasticity of the pine processional moth (Thaumetopoea pityocampa) populations under laboratory conditions. Entomologia, 3(1), 273. https://doi.org/10.4081/entomologia.2015.273

Berdan, E. L., Mazzoni, C. J., Waurick, I., Roehr, J. T., & Mayer, F. (2015). A population genomic scan in Chorthippus grasshoppers unveils previously unknown phenotypic divergence. Molecular Ecology, 24(15), 3918–3930. https://doi.org/10.1111/mec.13276

Branco, M., Paiva, M.-R., Santos, H., Burban, C., & Kerdelhué, C. (2017). Experimental evidence for heritable reproductive time in 2 allochronic populations of pine processionary moth, Insect Science, 24(2), 325–335. https://doi.org/10.1111/1744-7917.12287

Burbank, C., Gautier, M., Leblois, R., Landes, J., Santos, H., Paiva, M.-R., ... Kerdelhué, C. (2016). Evidence for low-level hybridization between two allochronic populations of the pine processional moth Thaumetopoea pityocampa (Lepidoptera: Notodontidae). Biological Journal of the Linnean Society, 119(2), 311–328. https://doi.org/10.1111/bij.12829

Catchen, J., Amores, A., Hohenlohe, P. A., Cresko, W., & Postlethwait, J. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1, 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. Molecular Ecology, 22(11), 3124–3140. https://doi.org/10.1111/mec.12354

Chessel, D., Dufour, A.-B., & Thioulouse, J. (2004). The ade4 package - I: One-table methods. R News, 4(1), 5–10.

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

Davey, J. W., & Blaxter, M. L. (2011). RADseq: Next-generation population genetics. Briefings in Functional Genomics, 9(5), 416–423. https://doi.org/10.1093/bfgp/elq031

Denlinger, D. L. (2002). Regulation of diapause. Annual Review of Entomology, 47, 93–122. https://doi.org/10.1146/annurev.ento.47.091201.145137

Derkarabetian, S., Burns, M., Starrett, J., & Hedin, M. (2016). Population genomic evidence for multiple Pliocene refugia in a montane-restricted harvestman (Arachnida, Opiliones, Sclerobunus robustus) from the southwestern United States. Molecular Ecology, 25(18), 4611–4631. https://doi.org/10.1111/mec.13789

Deks, M. F. L., Smit, S., Salis, L., Schijlen, E., Bossers, A., Mateman, C., ... Megens, H.-J. (2015). The genome of winter moth (Operophtera brumata) provides a genomic perspective on sexual dimorphism and phenology. Genome Biology and Evolution, 7(8), 2321–2332. https://doi.org/10.1093/gbe/evv145

Devaux, C., & Lande, R. (2008). Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and
